Modelos probabilísticos de tópicos para aplicaciones en Minería de Datos

CIMAT

Wilberth Ricardo García Alfaro
Departamento de Ciencias de la Computación
Centro de Investigación en Matemáticas, A.C.
Asesor: Dr. Salvador Ruiz Correa

Para la obtención del Grado de
Maestro en Ciencias de la Computación
18 de Noviembre de 2010

Dedicatoria

Quiero dedicar esta tesis a mis padres, los cuales siempre me han apoyado incondicionalmente, en todo momento y bajo toda circunstancia. A ustedes que han sabido ser los mejores padres, les debo todo lo que sé y lo que soy.

A ti madre, que siempre me impulsas a seguir adelante, me comprendes y apoyas en mis decisiones, que no me juzgas a pesar de lo imperfecto que soy, pero siempre me has corregido y enseñando a aprender de mis errores.

A ti padre que me has enseñado, inculcándome valores, a terminar bien todo lo que se empieza, con tu ejemplo me has mostrado el valor del trabajo honrado y me has educado toda la vida.

También quiero dedicar este documento a Anshela, por tu paciencia, comprensión, pero sobre todo por tu gran amor. Esta espera ha sido larga y difícil, pero a tu lado mi vida siempre es mejor, me haces sentir feliz y ser una mejor persona. Porque la distancia solo pudo separarnos físicamente, pero junto más nuestros corazones; porque siempre supiste convertir lo adverso en algo positivo; porque compartes conmigo planes y sueños, porque comparto la visión de un futuro juntos; y porque a pesar de todo a lo largo de estos años los recuerdos de nosotros siempre han sido los mejores.

A ustedes tres que siempre han sabido ser parte importante de mi vida, que hemos pasado penas juntos, que hemos festejado éxitos parciales y sufrido angustias a lo largo de este proceso, no solo les dedico esta tesis, sino también les dedico la maestría, porque parte de haber alcanzado esta meta, ha sido gracias a todo lo que me han dado, enseñado y compartido, por su sabiduría en forma de consejos, sus interminables buenos deseos, oraciones y el amor que han sabido transmitir, a pesar de la distancia. Nunca olvidare todo lo que hemos pasado juntos.

¡Gracias!

Agradecimientos

En primer lugar agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el financiamiento otorgado a lo largo de mis estudios de posgrado.

Quisiera agradecer a mi amigo y asesor Salvador, por permitirme trabajar bajo su supervisión, y darme la oportunidad de aprender de él no solo en lo académico, sino también en lo personal. Ha sido una experiencia muy enriquecedora en muchos sentidos.

También quisiera agradecer a todos mis compañeros de la maestría, ya que siempre nos hemos apoyado y ayudado cuando las circunstancias lo ameritaban y juntos compartimos todo el proceso de aprendizaje que involucró la maestría.

Resumen

Los modelos estadísticos establecen la relación entre las variables aleatorias involucradas en un proceso. Mediante el uso de parámetros hacen posible estudiar la estructura de una colección de datos desde un punto de vista matemático. Se auxilia con herramientas tales como el muestreo o la inferencia, y mediante un conjunto de suposiciones acerca de las distribuciones de la población, intenta explicar el comportamiento de dicho conjunto.

El primer paso para sugerir un modelo estadístico, es considerar la naturaleza de los datos. Esto ocasiona que se hagan suposiciones acerca de la forma en la cual se generan las muestras a partir de la población. Posteriormente, se proponen métodos eficientes para encontrar los parámetros de los modelos.

Como su nombre lo indica esta tésis describe el uso de dos de los modelos estadísticos llamados "modelos de tópicos" usados como técnicas de minería de datos.

Los modelos de tópicos intentan representar la estructura de los datos mediante el uso de distribuciones multinomiales, a partir de las cuales, se asigna una mayor probabilidad de aparición de los valores en la colección. Estas distribuciones de probabilidad son llamadas tópicos.

Esta familia de modelos, nace inicialmente en el área de minería de textos, pero actualmente es utilizada en otros contextos. En este documento se presentan y desarrollan los modelos conocidos como "LDA" (Latent Dirichlet Allocation) y el de "Autores y Tópicos".

La tésis está organizada de la siguiente manera:

- Introducción. En este capítulo se explicará de forma breve, la motivación detrás del uso de los modelos de tópicos para hacer minería de datos. Además, se discutirá el funcionamiento básico de estos modelos.
- Modelos de gráficas probabilísticas. Este capítulo forma parte de los conceptos básicos para desarrollar la teoría detrás de los

modelos LDA y de Autores y Tópicos. Ambos casos, son introducidos en forma de modelos de gráficas probabilísticas.

- Markov chain Monte Carlo (MCMC). Como se menciono con anterioridad, los modelos de tópicos aquí discutidos, son resueltos a través de simulaciones Monte Carlo. Por esta razón, en este capítulo se expone una breve explicación de esta familia de métodos, así como la forma en que funciona.
- Latent Dirichlet Allocation (LDA). En este capítulo se presenta la forma en la que se obtuvieron las ecuaciones que permiten resolver el modelo. También se ejemplifican sus aplicaciones mediante el uso de diversas bases de datos.
- Modelo de Autores y Tópicos. De nueva cuenta se introduce y desarrolla este modelo, el cual resulta ser una extensión del LDA. Se muestran aplicaciones y ejemplos relevantes y los resultados obtenidos de los experimentos realizados.
- Discución. Se desarrollan algunos comentarios acerca de las ventajas y desventajas del uso de los modelos LDA y de Autores y Tópicos. También se discute acerca de las mejoras y costes computacionales que representan las diversas implementaciones de los algormitmos, así como, las circunstancias en las que son recomendables.
- Conclusiones. Se exponen las deducciones obtenidas de los experimentos para ambos modelos.

Una descripción rápida de los principios en los que se basa el funcionamiento de los Métodos Variacionales, como lo es el Método Variacional Bayesiano, se presenta en el anexo A. En el anexo B se incorporan algunos documentos usados en los experimentos del capítulo "Latent Dirichlet Allocation" que apoyan y confirman los resultados. Para un mejor datalle de los resultados relacionados con autores y documentos obtenidos con el Modelo de Autores y Tópicos consulte el anexo C. El último anexo D es una descripción general del software desarrollado.

Índice general

Nomenclature **x**

1. Introducción **1**

2. Modelos de Gráficas Probabilísticas **4**
2.1. Redes Bayesianas 6
2.2. Independencia Condicional 11
2.3. Separación D 14
2.4. Cobija de Markov 16

3. Markov chain Monte Carlo (MCMC) **18**
3.1. Cadenas de Markov 20
3.2. Integración Monte Carlo 22
3.3. Muestreo de Gibbs 22
3.4. Dificultades para hacer inferencia usando el MCMC 24
3.4.1. Monitoreo de convergencia y problemas ocasionados por la convergencia lenta 25
3.4.2. Estudio de la autocorrelación 27
3.4.3. Gráficas de Kernel 28
3.4.4. Prueba Z de Geweke 30
3.4.5. Método de Gelman-Rubin 31

4. Latent Dirichlet Allocation (LDA) **33**
4.1. Modelo de gráficas probabilísticas y proceso generativo del LDA . 34
4.2. LDA suavizado 35
4.3. Muestreo de Gibbs colapsado para LDA suavizado 42
4.3.1. Valores esperados de los parámetros 45
4.4. Semántica y LDA 46
4.5. Aplicaciones del LDA 47
4.5.1. Similitud entre documentos y entre palabras 47
4.5.1.1. Análisis de similitud entre documentos 47

4.5.1.2. Similitud entre palabras 48
4.5.2. Aplicaciones en recuperación de información 49
4.5.3. Agrupamiento de Documentos 49
4.5.4. Análisis de la tendencia entre los tópicos 50
4.5.4.1. Tópicos de moda 50
4.5.4.2. Progresión de tópicos de interés por períodos de tiempo 50
4.6. Experimentos 50
4.6.1. Experimento sintético 1 51
4.6.2. Experimento sintético 2 55
4.6.2.1. Selección de modelo 58
4.6.3. Experimento de texto con datos reales 61
4.6.3.1. Base de datos de NIPS 61
4.6.4. Base de datos de WormsBase 64

5. Modelo de Autores y Tópicos 86
5.1. Modelo de Gráficas y Proceso Generativo 87
5.2. Muestreo de Gibbs para el Modelo de Autores y Tópicos 88
5.3. Aplicaciones del Modelo de Autores y Tópicos 97
5.3.1. Análisis de Tendencia de los Autores por Año 98
5.3.2. Detección de Documentos poco Comunes 98
5.3.3. Comparación de Tópicos entre Autores 99
5.3.4. Etiquetado Automático de Nuevos Documentos para Autores en la Colección 99
5.4. Experimentos 100
5.4.1. Experimento con sintético 100
5.4.2. Base de datos de NIPS 106
5.4.3. Base de datos de WormBase 112

6. Discución 123
6.1. Beneficios del Uso del Muestreo de Gibbs 124
6.2. Beneficios de Usar un Modelo 125

7. Conclusiones 129
7.1. Trabajo a futuro 130

A. Métodos Variacionales 131

B. Recopilación de resúmenes para clasificación y agrupamiento 135
B.1. Agrupamiento 135
B.1.1. Grupo 1 136
B.1.2. Grupo 6 140

B.1.3. Grupo 11 144
B.1.4. Grupo 29 148
B.2. Recuperación 152

C. Recopilación de datos para el Modelo de Autores y Tópicos 158
C.1. Títulos por autor para NIPS 158
C.1.1. Tópico 6 159
C.1.2. Tópico 16 162
C.1.3. Tópico 31 168
C.1.4. Tópico 44 170
C.1.5. Tópico 50 173
C.2. Títulos por autor para wormbase 175
C.2.1. Tópico 3 175
C.2.2. Tópico 18 178
C.2.3. Tópico 29 181
C.2.4. Tópico 32 184
C.2.5. Tópico 48 187
C.2.6. Tópico 50 190

D. Descripción del Software 193

Referencias 196

Índice de figuras

2.1. Modelo de gráficas probabilísticas de la ecuación 2.3. 7
2.2. Modelo de gráficas probabilísticas con múltiples variables. 8
2.3. Modelo de gráficas probabilísticas en *Plate Notation* equivalente al de la figura 2.2. 9
2.4. Modelo de gráficas probabilísticas de la regresión polinomial considerando parámetros. 9
2.5. Modelo de gráficas probabilísticas de la regresión polinomial considerando parámetros y con variables observadas. 10
2.6. Modelo de gráficas conocido como "divergente". 12
2.7. Modelo de gráficas conocido como "serial". 13
2.8. Modelo de gráficas conocido como "convergente". 14
2.9. Modelo de gráficas de una muestra i.i.d. para el caso de una gaussiana univariada. 15
2.10. Un ejemplo de la Cobija de Markov para la variable x_i. 17

3.1. Evolución de dos medias en el proceso de muestreo de la cadena de Markov. 21
3.2. Evolución de los parámetros θ_1 y θ_2 inicializados con diferentes valores. 26
3.3. Una aparente convergencia del parámetro. 27
3.4. Gráfica de la autocorrelación para los parámetros θ_1 y θ_2 con una correlación aún elevada. 28
3.5. Gráfica de la autocorrelación para los parámetros θ_1 y θ_2 con una correlación moderada. 28
3.6. Estimación de kernels que aún no demuestran convergencia. . . . 29
3.7. Estimación de kernels que demuestran convergencia. 30

4.1. Modelo generativo probabilístico LDA. 34
4.2. Modelo generativo probabilístico LDA suavizado. 36
4.3. Ejemplo de la palabra "banco" manejada en diferentes contextos. . 46
4.4. Tópicos desde los cuales fueron generadas los patrones de franjas. 55

4.5. Un subconjunto de datos de entrenamiento usados para aprender los parámetros del modelo. 56
4.6. Evolución de los tópicos en diferentes iteraciones. 57
4.7. Evolución de la estimación con respecto al número de iteraciones. 57
4.8. Diversos valores de β y $\alpha = 1$ para la gráfica de la selección de modelo. 60
4.9. Abstract del artículo titulado "Two Iterative Algorithms for Computing the Singular Value Decomposition from Input/Output Samples" escrito por Terence D. Sanger y obtenido de la base de datos de NIPS. 63
4.10. Proporción de aparición de los tópicos en una muestra aleatoria de 100 documentos de la colección. 66
4.11. Mezcla de tópicos para el documento titulado "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin". 66
4.12. Extracto del resumen del artículo titulado "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin" etiquetados de forma automática. 67
4.13. Proporción de aparición de la primera palabra de cada tópico. . . 68
4.14. Distribución de los tópicos para los documentos consulta, el más parecido y el menos parecido. 70
4.15. Parte del resumen perteneciente al documento 55. 71
4.16. Parte del abstrac perteneciente al documento 31. 71
4.17. Dendograma de la muestra de 100 documentos presentados en la figura 4.10. 72
4.18. Distribución de los tópicos para los documentos del grupo 1. . . . 75
4.19. Distribución de los tópicos para los documentos del grupo 6. . . . 78
4.20. Distribución de los tópicos para los documentos del grupo 11. . . 80
4.21. Distribución de los tópicos para los documentos del grupo 29. . . 83
4.22. Gráfica de $p(q|d_i)$ para la recuperación de documentos. 84

5.1. Modelo de gráficas probabilísticas del Modelo de Autores y Tópicos. 87
5.2. Modelo de Autores y Tópicos extendido. 96
5.3. Tópicos desde los cuales fueron generadas las imágenes. 100
5.4. Tópicos obtenidos al correr el algoritmo para el modelos de Autores y Tópicos. 101
5.5. Gráfica de la perplejidad aplicada a los resultados obtenidos mediante el muestro de Gibbs para LDA, Variational Bayes y muestreo de Gibbs para el Modelo de Autores y Tópicos. 102
5.6. Distribución de los autores y los tópicos. 103

5.7. Análisis de evolución para tres autores por año. 117
5.8. Resumen perteneciente el documento titulado "slo-1 modulation of neuronal activity in the pharynx". 118
5.9. Resumen perteneciente el documento titulado "Involvement of aak-2 and insulin-like signalling mutations in the cellular stress response as determined by an in vivo ATP sensor C.elegans strain". . . . 119

B.1. Parte del resumen perteneciente al documento 1. 136
B.2. Parte del resumen perteneciente al documento 79. 137
B.3. Parte del resumen perteneciente al documento 85. 138
B.4. Parte del resumen perteneciente al documento 95. 139
B.5. Parte del resumen perteneciente al documento 6. 140
B.6. Parte del resumen perteneciente al documento 18. 141
B.7. Parte del resumen perteneciente al documento 23. 142
B.8. Parte del resumen perteneciente al documento 82. 143
B.9. Resumen perteneciente el documento 11. 144
B.10.Parte del resumen perteneciente al documento 28. 145
B.11.Parte del resumen perteneciente al documento 34. 146
B.12.Parte del resumen perteneciente al documento 93. 147
B.13.Parte del resumen perteneciente al documento 29. 148
B.14.Parte del resumen perteneciente al documento 60. 149
B.15.Parte del resumen perteneciente al documento 66. 150
B.16.Resumen perteneciente al documento 91. 151
B.17.Parte del resumen perteneciente al documento 19. 153
B.18.Resumen perteneciente al documento 26. 154
B.19.Parte del resumen perteneciente al documento 55. 155
B.20.Parte del resumen perteneciente al documento 34. 156
B.21.Parte del resumen perteneciente al documento 93. 157

Capítulo 1

Introducción

Cuando se habla de minería de datos, se habla de un conjunto de técnicas, algoritmos y procesos que permiten extraer información valiosa acerca de los datos que se estudian.

Por esta razón es muy importante conocer la diferencia que existe entre los términos: datos, información y conocimiento. Los datos son un conjunto de representaciones simbólicas de un atributo o característica que tiene una entidad, cuya principal propiedad es carecer de sentido propiamente. Cuando un conjunto de datos es procesado en algún determinado contexto, estos adquieren relevancia para un propósito convirtiéndose en información. Por otro lado, el término conocimiento se atribuye a individuos que han procesado un conjunto de datos y de información, tratando de aplicarlos para la toma de decisiones con la finalidad de resolver un problema en específico. En resumen, se puede pensar en los datos, como la información cruda antes de ser procesada mientras que el conocimiento es la articulación de la información para un propósito específico.

El interés de generar conocimiento, se encuentra en tratar de mejorar el rendimiento en la resolución de problemas. Por este motivo una colección de datos, no es útil si en algún momento no es procesada para generar información, y dicha información es usada para generar conocimiento. Sin esta progresión de eventos una colección de datos simplemente permanecería ociosa sin ningún uso ni provecho.

Pero en la actualidad el desarrollo de la tecnología, se ha visto acompañado por el aumento de la capacidad de almacenamiento y por lo tanto han proliferado colecciones de datos para diversos usos. Asimismo, las bases de datos han crecido en tamaño, por lo tanto el procesamiento del cual se obtiene información valiosa, ya no es más una tarea trivial.

Para poder resolver el problema de procesar una gran cantidad de datos, se han desarrollado una serie de herramientas, en su mayoría construidas sobre la base de la probabilidad y la estadística. Estas permiten extraer información útil, que se encuentra de manera inherente en los datos.

La metodología de trabajo en el área de minería es extraer esta información de tal manera que pueda ser expresada de forma comprensible y resumida, encontrando los patrones que sintetizan la relación entre los datos, así como su estructura básica.

Es por esto, que en algunos autores como en Witten y Frank (2005), definen minería de datos como el proceso mediante el cual se descubren patrones útiles de forma automática o semiautomática, en grandes cantidades de datos. La utilidad de estos patrones radica en la posibilidad de mejorar la compresión de la estructura subyacente de los datos, para poder responder preguntas acerca del porqué ocurren los comportamientos o circunstancias estudiados, poder predecir nuevos resultados cuando se tiene cierta información.

Una vez obtenida la estructura de los datos, se suele resumir la información a través de diversos tipos de estrategias, como son las reglas de decisión, árboles de decisión, grupos para la clasificación y modelos matemáticos o estadísticos que permiten expresar de forma más entendible la relación entre los datos que se estudian.

En particular el trabajar con modelos estadísticos, proporciona un entendimiento extra acerca de la información contenida. Esto ocurre típicamente debido a que con los modelos matemáticos y estadísticos es posible implementar diversas herramientas para extraer información desde diferentes enfoques, y en ocasiones capturan de forma más completa y eficiente la estructura de la información.

Debido a estas ventajas, en este trabajo se discuten algunos temas de minería de datos, usando como estrategia la estimación de modelos estadísticos para la extracción de información en colecciones de texto.

Un modelo estadístico en términos simples, es un conjunto de ecuaciones matemáticas, que a través del uso de distribuciones de probabilidad, establecen la estructura que existe en los datos. Los modelos estadísticos, hacen uso de las observaciones, y a través de las técnicas de inferencia estadística, tratan de encontrar los valores adecuados de los parámetros que involucran las distribuciones de probabilidad establecidas por el modelo y que generalizan a la población de estudio, únicamente a partir de las muestras.

El modelo estadístico a través del uso de las distribuciones de probabilidad, en muchas ocasiones hace suposiciones acerca de la naturaleza de la estructura de los datos. Al establecer el número de variables aleatorias involucradas en el modelo, así como la forma en la que se distribuye la población y el valor de los parámetros, los modelos estadísticos ya presuponen muchas características de los datos como la forma en la que fueron creados, las propiedades de la población de donde provienen e inclusive el significado de cada valor en el contexto de la base de datos.

Los modelos de tópicos son una familia de modelos estadísticos que presuponen que cada símbolo dentro de la base de datos fue generado por la aparición de

un determinado valor en una *variable latente (porque no es observada)*, que a su vez está asociado con un tópico. En esencia, un tópico es una distribución de probabilidad que asigna a cada valor del conjunto de datos observado una probabilidad de aparición. Este tipo de modelos, cuentan además con un algoritmo conocido como proceso generativo, que representa el proceso de muestreo que debe de seguir para obtener una muestra de la población similar al conjunto de datos que se estudian.

Los modelos de tópicos surgieron del área de minería de textos, donde se han modelado los tópicos de los textos con los que se escriben los documentos, mediante el uso de palabras relacionadas estadísticamente. La aparición de palabras es modelada a través de distribuciones multinomiales de probabilidad y estas distribuciones relacionan las palabras de tal manera que reflejan un contenido semántico, contenido que caracteriza a cada uno de los tópicos. A pesar de que de forma natural la semántica se relaciona con cuestiones de lenguaje natural (el mejor ejemplo del manejo de lenguaje natural se encuentra en los textos), se han podido aplicar este tipo de modelos para modelar otras situaciones en donde la idea de agrupar palabras con tópicos también es válida.

La idea de usar estos modelos de tópicos, está basada en el hecho de que si los datos fueron observados, entonces estos son los que tienen la mayor probabilidad de aparición, y si se cuentan con suficientes muestras, se logra representar de forma adecuada las características de la población. De esta manera al encontrar los parámetros que explican de mejor la aparición de los datos observados, es posible generalizar dichos resultados a toda la población de donde provienen nuestras muestras.

Cada modelo de tópico se resuelve de forma diferente, dependiendo de las asunciones que se hagan. Sin embargo, en ocasiones algunos modelos tópicos son extensiones de otros ya conocidos, que tratan de adaptarlo a diferentes situaciones en el análisis de las bases de datos. En este documento se estudian dos modelos de tópicos conocidos. En el capítulo 4 se explica y desarrolla el modelo "Latent Dirichlet allocation" o LDA, mientras que en el capítulo 5 se discute el modelo conocido como de "Autores y Tópicos", el cual es una extensión del modelo LDA que asigna una probabilidad de responsabilidad a cada autor para todos los documentos de la colección, con base en la probabilidad de aparición de los tópicos para cada autor.

Capítulo 2

Modelos de Gráficas Probabilísticas

Desde que la construcción de las colecciones de datos involucra procesos muy complejos, el hacer minería en ellas resulta también en una tarea compleja. Los modelos de gráficas probabilísticas forman un marco de trabajo muy extenso y bien estudiado. Estos permiten abordar el problema de realizar inferencia probabilística y estadística en un conjunto de datos, con el objeto de encontrar patrones que sintetizen la relación entre ellos. De esta manera, los modelos de gráficas probabilísticas utilizados en el contexto de minería de datos tienen una doble utilidad:

1. Ayudan a comprender el proceso detrás de la generación de los datos observados.

2. Proporcionan una forma eficiente y tratable de extraer la información deseada.

El primer punto es realizado cuando un modelo sugiere el proceso que generó el conjunto de datos observado.

La segunda característica es alcanzada al transformar los datos de cualquier índole, en información estadística que permita la aplicación de diversos análisis que faciliten la obtención de información concreta y útil.

Debido a que los modelos de tópicos son en esencia modelos de gráficas probabilísticas, la teoría expuesta en este capítulo resulta ser fundamental para la comprensión de la forma en qué estos operan y los pasos a seguir para la obtención de las ecuaciones que eventualmente llevan a resolver el problema de hacer minería en colecciones de textos.

Un modelo de gráficas probabilísticas es en esencia una gráfica que representa a una familia de distribuciones de probabilidad. De acuerdo con Jordan (1999), los modelos de gráficas probabilísticas forman parte de un área del modelado matemático que está íntimamente relacionada con la teoría de la probabilidad

y la teoría de las gráficas. Proporcionan una herramienta natural que permiten tratar con dos problemas que aparecen con suma frecuencia en los campos de matemáticas aplicadas e ingeniería: la incertidumbre y la complejidad.

La idea fundamental detrás de un modelo de gráficas es la noción de modularidad. Es decir, ver un sistema complejo como una combinación de partes más simples. Por otro lado, la teoría de la probabilidad proporciona una forma de unir dichas partes, asegurándose de que el sistema visto como un todo, es consistente y provee una forma de interacción entre el modelo y los datos.

En otras palabras, los modelos de gráficas probabilísticas permiten crear representaciones intuitivamente más atractivas. Con ellos, las personas pueden modelar conjuntos de variables aleatorias cuyas interacciones son muy complejas, conservando la estructura de los datos que tienen de forma intrínseca. Esta simplificación, se realiza con la finalidad de diseñar algoritmos eficientes de propósito general.

Según Bishop (2007) algunas de las ventajas de usar modelos de gráficas probabilísticas en la práctica son:

1. Proporcionan una forma muy simple de visualizar la estructura de los modelos probabilísticos que puede ser usados para el diseño o motivación de nuevos modelos.

2. Facilitan el análisis de las propiedades del modelo, incluidas las propiedades de independencia condicional que pueden ser obtenidas por simple inspección.

3. Facilitan la ejecución de cálculos complejos requeridos para procesos de inferencia y aprendizaje en modelos muy sofisticados. Dichos cálculos pueden ser expresados en términos de manipulaciones gráficas en las cuales las expresiones matemáticas son realizadas implícitamente.

Una gráfica está formada por un conjunto de nodos o vértices que en un modelo de gráficas probabilísticas representa cada una de las variables aleatorias o grupos de variables aleatorias; y los lados o aristas representan las relaciones probabilísticas entre las variables de las cuales se forma. En esencia, una gráfica captura la forma en que cada distribución conjunta de probabilidad sobre todas las variables se descompone en un producto de factores, cada uno dependiente de un subconjunto de esas variables. Esta estrategia permite una reducción en la complejidad de los cálculos al trabajar únicamente con conjuntos de variables aleatorias en lugar de trabajar directamente con todas.

Ahora bien, como es sabido las gráficas de dividen en dirigidas y no dirigidas. En los modelos de gráficas dirigidas la dirección de las aristas establecen el orden de la relación; este tipo de modelos es conocido como **Redes Bayesianas**.

Mientras que las gráficas no dirigidas son conocidas como **Campos Aleatorios de Markov**, en los cuales no existe un orden específico en la relación de dichas variables.

2.1. Redes Bayesianas

Esta sección intentará proporcionar una motivación del porque usar modelos de gráficas probabilísticas para representar distribuciones de probabilidad.

Primeramente, recuerde las dos propiedades fundamentales de la teoría de la probabilidad, las cuales se enuncian a continuación:

$$p(X) = \sum_{Y} p(X, Y) \tag{2.1}$$

$$p(X, Y) = p(Y|X)P(X) \tag{2.2}$$

La ecuación 2.1 es conocida como **la regla de la suma**, mientras que la ecuación 2.2 es conocida como **la regla del producto**. Estas dos reglas son la base fundamental de la teoría de la probabilidad.

Ahora bien, considere la distribución de probabilidad conjunta $p(a, b, c)$ sobre las variables a, b, c. Utilizando la regla del producto 2 veces se puede escribir esta probabilidad como:

$$p(a, b, c) = p(c|b, a)p(a.b) = p(c|b, a)p(b|a)p(a) \tag{2.3}$$

Para representar esta distribución de probabilidad como un modelo de gráficas probabilísticas basta tomar cada distribución de probabilidad condicional y poner cada variable aleatoria dentro de un nodo. Posteriormente, crear una arista dirigida desde los nodos con las variables aleatorias que condicionan hacia las variables no condicionantes. Por ejemplo, la arista que va de a hacia b representa el término $p(b|a)$. La representación del modelo gráfico de la ecuación 2.3 se muestra en la figura 2.1.

Se debe señalar que en este modelo de gráficas probabilísticas no se proporciona información acerca del dominio de las variables (si son continuas o discretas), ni tampoco la distribución de probabilidad marginal de ellas, por lo que en realidad el modelo de gráficas como se ha mencionado anteriormente, representa no únicamente una distribución de probabilidad sino a todas aquellas que se puedan factorizar de la forma mostrada en la ecuación 2.3. Otra cuestión importante de

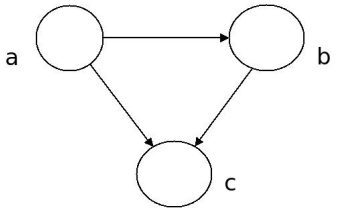


Figura 2.1: Modelo de gráficas probabilísticas de la ecuación 2.3.

señalar, es que el tomar un orden de factorización diferente al tomado en la ecuación 2.3 genera un modelo de gráficas diferente pero equivalente al mostrado en la figura 2.1.

En general, se puede extender la idea de factorizar la Red Bayesiana como un producto de probabilidades condicionales de los nodos dados sus antecesores (también conocidos como padres) y aplicando de forma repetitiva la regla del producto se tiene que:

$$p(x_1, x_2, ..., x_K) = p(x_K|x_1, ..., x_{K-1})...p(x_2|x_1)p(x_1) \tag{2.4}$$

De la ecuación 2.4 se puede apreciar que para un conjunto de K variables siempre se puede escoger un orden, de tal manera que cada nodo aparezca como un factor con los nodos antecesores como condicionantes, así se puede expresar de forma un poco más general para una gráfica dirigida la ecuación 2.4 en términos de sus nodos antecesores o padres, es decir:

$$p(x_1, x_2, ..., x_K) = \prod_{k=1}^{K} p(x_k|pa(x_k)) \tag{2.5}$$

Donde $pa(x_k)$ representa a los nodos antecesores de la variable x_k, es decir, los nodos que tienen una arista hacia el nodo x_k.

Esta ecuación establece la relación entre una distribución de probabilidad sobre todas las variables y una gráfica dirigida, la relación también se cumple en el otro sentido.

Otra restricción de suma importancia para que la ecuación 2.5 sea válida, es que la gráfica dirigida sea acíclica (DAG por sus siglas en inglés), o sea, no de deben existir caminos cerrados que permitan ir de un nodo determinado en la dirección de las flechas y que finalice en el mismo nodo de incio. Esta restricción es la que permite afirmar que si existe un orden de factorización de dicha distribución de probabilidad entonces es posible generar un DAG y viceversa.

Como última observación se puede mencionar que a pesar de que en la ecuación 2.5 se hace referencia a cada variable x_k como una sola variable en realidad estas pueden representar vectores de variables u otros tipos de variables más complejas.

Ahora bien, considere el siguiente modelo de gráficas mostrado en la figura 2.2.

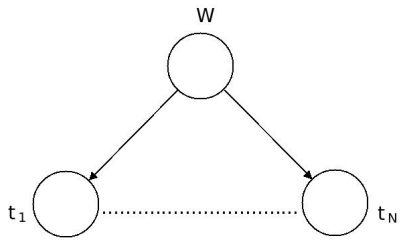


Figura 2.2: Modelo de gráficas probabilísticas con múltiples variables.

Este modelo de gráfica fue introducido por Bishop (2007) y es conocido con el nombre de Regresión Polinomial. La distribución de probabilidad conjunta del modelo presentado en la figura 2.2 está dada por:

$$p(\mathbf{w},\mathbf{t}) = p(\mathbf{w})\prod_{n=1}^{N} p(t_n|\mathbf{w}) \tag{2.6}$$

Donde $\mathbf{w}$ es un vector de coeficientes de un polinomio y $\mathbf{t} = (t_1, t_2, ..., t_N)^T$ es un vector de observaciones en el tiempo T.

En ocasiones cuando se trabaja con procesos muy grandes, la aparición de ciclos puede producir un modelo de gráficas muy extenso y cuya visualización es difícil.

Para solventar este problema se introduce una nueva notación llamada *Plate Notation* en la cual basta sustituir los ciclos por submodelos de gráficas que representan el proceso llevado a cabo en el interior del ciclo. Para denotar la existencia de iteraciones, estos submodelos suelen ir encerrados en un rectángulo en cuya esquina inferior derecha se anota el número de veces que se repite el proceso. También es común el uso de subíndices, para indicar que las variables no son variables simples, sino un conjunto.

En el caso de la figura 2.2 basta con encerrar un nodo que representa a la variable t_n en un rectángulo y colocar en la esquina la N indicando que el subíndice n corre desde 1 hasta N. La figura 2.2 queda de una forma más compacta de la forma en la que se visualiza en la figura 2.3

Sin embargo, en el modelo de la figura 2.2 a diferencia del modelo original de regresión polinomial, no considera un conjunto de parámetros como lo son:

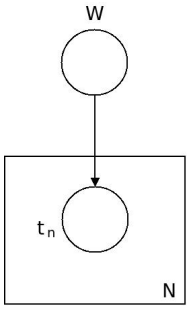


Figura 2.3: Modelo de gráficas probabilísticas en *Plate Notation* equivalente al de la figura 2.2.

.

el conjunto de datos de entradas para las variables observadas, un término de ruido que se denota por σ^2 (que también afecta a las variables observadas) y finalmente un parámetro de selección de la gaussiana que afecta a la variable W. La distribución de probabilidad considerando los parámetros se muestra en la ecuación 2.7.

$$p(\mathbf{w},\mathbf{t}|\mathbf{x},\sigma^2,\alpha) = p(\mathbf{w}|\alpha)\prod_{n=1}^{N} p(t_n|\mathbf{w},x_n,\sigma^2) \tag{2.7}$$

El modelo de gráficas de la regresión polinomial ya con los parámetros queda como se muestra en la figura 2.4:

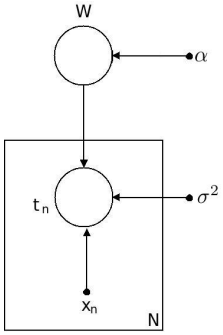


Figura 2.4: Modelo de gráficas probabilísticas de la regresión polinomial considerando parámetros.

.

Los parámetros en el modelo de gráficas están representados por un pequeño círculo al comienzo de la arista y las variables a un lado de la arista sin estar encerradas en ningún círculo hacen referencia a los hiperparámetros. La punta de

la flecha debe estar dirigida hacia la variable a la cual condiciona. Las reglas de la notación en *Plate Notation* se aplica de la misma manera.

Como se señalo anteriormente, las variables t_n hacen referencia a los valores observados de la salida del polinomio, por lo que también es importante destacar en el modelo de gráficas que dichas variables son observadas. Para esto se rellena el círculo de la variable aleatoria indicando que esa variable o grupo de variables son observadas. El modelo de gráficas anterior queda entonces como:

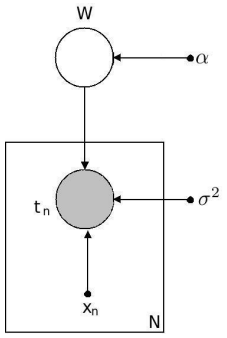


Figura 2.5: Modelo de gráficas probabilísticas de la regresión polinomial considerando parámetros y con variables observadas.

.

Las variables no observadas en el modelo de la figura 2.5 son llamadas variables latentes o escondidas y son el objeto de estudio de muchos métodos estadísticos y probabilísticos usados en el área de aprendizaje por computadora. Ya que estas determinan el valor de las variables observadas, pueden ser interpretadas como causalidad. Es por esta razón, por la que muchos modelos probabilísticos basan su funcionamiento en el cálculo de probabilidad posterior (la probabilidad de los parámetros dados los datos observados).

Debido a esta característica de causalidad, típicamente en un modelo gráfico se pueden ver a las variables latentes como las variables que condicionan y que son padres o ancestros de las variables observadas. Por esta razón, las numeraciones más altas suelen ser ocupadas por las variables observadas.

En resumen, los modelos de gráficas tratan de explicar las relaciones entre las variables. También señalan la forma en el cual cierto conjunto de valores de las variables latentes dan lugar al conjunto de datos observados en las variables observadas, capturando la causalidad. Es por esto, que en ocasiones estos modelos son conocidos con el nombre de modelos generativos.

2.2. Independencia Condicional

Considere ahora un conjunto de tres variables aleatorias a, b, c. Se dice que a es condicionalmente independiente de b si se cumple para cada valor que toma la variable c que :

$$p(a|b,c) = p(a|c) \tag{2.8}$$

Y se denotará de forma más compacta como:

$$a \perp\!\!\!\perp b \,|\, c \tag{2.9}$$

Otra definición alternativa para independencia condicional esta dada por:

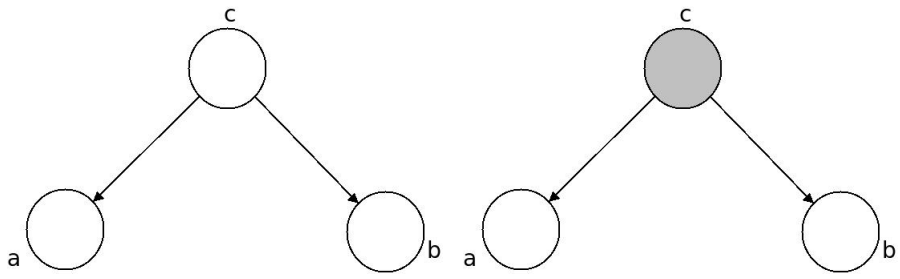
$$p(a,b|c) = p(a|c)p(b|c) \tag{2.10}$$

Las propiedades de independencia condicional simplifican los cálculos al trabajar con un conjunto más pequeño de variables. Cuando se trabaja con la forma algebraica de una distribución de probabilidad, encontrar las propiedades de independencia condicional involucra un gran número de operaciones haciendo muy compleja dicha tarea. Utilizando el modelo de gráficas estas propiedades pueden ser obtenidas con solo analizar el DAG a través del criterio de "separación d" inicialmente propuesto en Pearl (1988) y cuya prueba puede ser consultada en Lauritzen (1996).

Para comenzar a explicar las propiedades de independencia condicional se presentan tres ejemplos, el primero de los cuales se desarrolla a continuación. Considere el siguiente modelo gráfico mostrado en la figura 2.6a. Este modelo debido a la notación antes introducida tiene la distribución conjunta de probabilidad mostrada en la ecuación 2.11. Si se marginaliza la variable c en la ecuación 2.10 se aprecia que se obtiene el producto de $p(a)$ y $p(b)$. De igual forma, para verificar las propiedades de independencia condicional del modelo 2.6a se procede a marginalizar con respecto a c, como se muestra en la ecuación 2.12.

$$p(a,b,c) = p(a|c)p(b|c)p(c) \tag{2.11}$$

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \tag{2.12}$$



(a) Sin variables observadas. **(b)** Se observa la variable c.

Figura 2.6: Modelo de gráficas conocido como "divergente".

Se puede apreciar que la ecuación 2.12 no factoriza en el producto esperado, por lo que en el modelo de gráficas divergente de la figura 2.6a se puede afirmar que $a \not\perp\!\!\!\perp b \mid \emptyset$ o bien solo $a \not\perp\!\!\!\perp b$.

Por el contrario, el modelo de gráficas de la figura 2.6b según la notación se factoriza en la ecuación 2.13, demostrando que $a \perp\!\!\!\perp b \mid c$.

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a|c)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned} \tag{2.13}$$

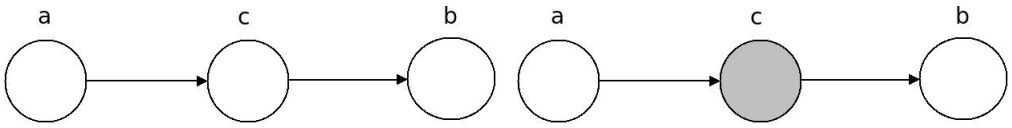
Una forma de interpretar de gráficamente este resultado ocurre cuando se considera una ruta de recorrido de a hacia b. Dicha ruta forzosamente pasa por el nodo c en cual recibe la parte de atrás de la flecha.

Cuando un nodo tiene esta disposición se dice que es un nodo "cola a cola", ya que en inglés se denomina la tail (traducción de cola)a la parte que no es la punta de la flecha.

Debido a que la única ruta existente entre a y b pasa por c, y dado que el nodo c no está observado, entonces, a y b se vuelven dependientes. Por otro lado, si el valor de c es observado entonces dicho camino está bloqueado causando la independencia de las variables.

De nueva cuenta considere el modelo de gráficas de la figura 2.7a en cual es conocido como modelo serial.

Se puede escribir la factorización del modelo de la figura 2.6a como se muestra en la ecuación 2.7a. Marginalizando con respecto a la variable c se obtiene la expresión 2.15, que en general no factoriza como el producto de téminos con las variables a y b únicamente.



(a) Sin variables observadas. **(b)** Se observa la variable c.

Figura 2.7: Modelo de gráficas conocido como "serial".

$$p(a,b,c) = p(b|c)p(c|a)p(a) \tag{2.14}$$

$$p(a,b) = p(a)\sum_{c} p(b|c)p(c|a) \tag{2.15}$$

En contraste, la distribución de probabilidad de la figura 2.7b está expresada en la ecuación 2.16. Se aprecia que esta distribución cumple con la definición de independencia condicional, por lo que se afirma que $a \perp\!\!\!\perp b \mid c$.

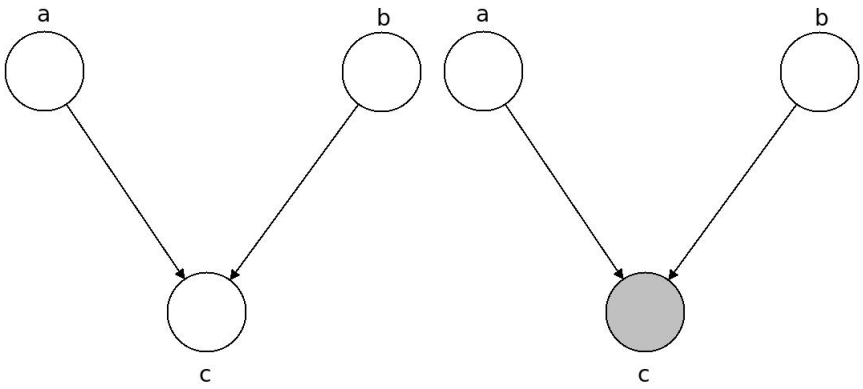
$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(b|c)p(c|a)p(a)}{p(c)} \\ &= \frac{p(b|c)p(a|c)p(c)p(a)}{p(c)p(a)} \\ &= p(b|c)p(a|c) \end{aligned} \tag{2.16}$$

Del mismo modo, es posible interpretar esta independencia verificando que el camino de a hacia b pasa forzosamente a través de c. La diferencia de los modelos de las figuras 2.7a y 2.7b, radica en el hecho de que la variable c se encuentra marcada como observada para el modelo 2.7b, y debido a la configuración "punta a cola" el único camino de a hacia b es bloqueado por c, logrando así la independencia condicional.

Finalmente, considere el último ejemplo llamado convergente que se presenta en la figura 2.8a.

La distribución de probabilidad del modelo de gráficas de la figura 2.8a se escribe como se muestra en la ecuación 2.17. Marginalizando con respecto a la variable c se verifica que $a \perp\!\!\!\perp b$ como se explica en la ecuación 2.18

$$p(a,b,c) = p(c|a,b)p(b)p(a) \tag{2.17}$$



(a) Sin variables observadas. **(b)** Se oberva la variable c.

Figura 2.8: Modelo de gráficas conocido como "convergente".

$$\begin{aligned} p(a,b) &= p(b)p(a)\sum_{c} p(c|a,b) \\ &= p(b)p(a) \end{aligned} \tag{2.18}$$

El modelo de gráficas de la figura 2.8b tiene como distribución de probabilidad conjunta la ecuación 2.19, la cual no factoriza como el producto de las variables a y b al marginalizar con respecto a c, por lo que $a \not\perp\!\!\!\perp b|\ c$.

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(c|a,b)p(b)p(a)}{p(c)} \end{aligned} \tag{2.19}$$

Esta configuración de nodos es conocida como "punta a punta". A diferencia de los modelos anteriores, el modelo convergente se bloquea el camino de a a b cuando el nodo c no es marcado como observado.

El comportamiento de la figura 2.8 es el que da lugar a un fenómeno conocido como *Explaining Away* cuyo dominio está fuera del alcance de este documento.

2.3. Separación D

En la sección anterior se mostraron tres de las estructuras básicas de los modelos de gráficas a partir de los cuales se forman modelos más complejos. También

se señalaron las motivaciones del porqué ocurren las propiedades de independencia condicional, así como, la relación que existe con el criterio de separación D. A continuación se enunciara de manera general dicho criterio.

Considere una gráfica dirigida y un grupo A, B, y C de variables tales que $A \cap B \cap C = \emptyset$ (cuya unión no es necesariamente el conjunto completo de vértices de la gráfica), si además, dicha gráfica es acíclica, se puede decir que $A \perp\!\!\!\perp B \mid C$ si todas rutas de A hacia B están bloqueadas.

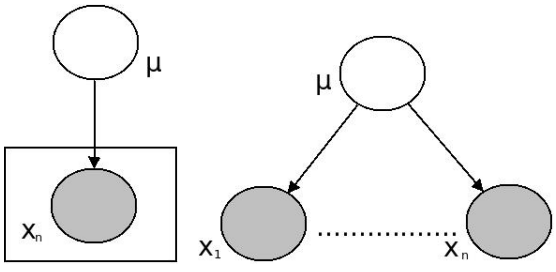
Se dice que una ruta está bloqueada si a lo largo del camino se encuentra un nodo, en el cual ocurre alguna de las siguientes situaciones:

a. Las aristas se unen "punta a cola" o "cola a cola" en el nodo y el nodo se encuentra en el conjunto C.

b. Las aristas se unen "punta a punta" y ni el nodo ni ninguno de los descendientes están en el conjunto C.

Si esto ocurre para todos las rutas que van de A hacia B, se puede concluir que el conjunto A está D-separado de B y por lo tanto todas la variables en los conjuntos A, B y C cumplen que $A \perp\!\!\!\perp B \mid C$.

Un ejemplo claro de este tipo de propiedades se puede ver en una muestra i.i.d. Considere el problema de estimar la distribución posterior de la media de una distribución gaussiana univariada. En la práctica uno observa los valores de las gaussianas y en base a esto se trata de estimar el valor adecuado de la media μ. El modelo de gráficas para esta aplicación se presenta en las figuras 2.9a y 2.9b.



(a) Modelo en *Plate Notation* . **(b)** Modelo extendido.

Figura 2.9: Modelo de gráficas de una muestra i.i.d. para el caso de una gaussiana univariada.

Ahora suponga que en dicho modelo se condiciona sobre μ y si se busca la distribución de probabilidad conjunta se obtiene la ecuación 2.20.

$$p(x|\mu) = \prod_{i}^{n} p(x_i|u) \tag{2.20}$$

Esta propiedad se verifica fácilmente usando la figura extendida 2.9b, en la cual, es notable que para ir de una variable x_i a otra $x_{j\neq i}$ se tiene que pasar a través de μ. Si este camino está bloqueado entonces se puede obtener que $x_i \perp\!\!\!\perp x_{j\neq i} \mid \mu$. Finalmente, la distribución conjunta se expresa como el producto de dichas variables, obteniendo la ecuación de 2.20.

2.4. Cobija de Markov

Considere un conjunto de variables $x_1, x_2, ...x_D$ de D variables. Ahora bien, si es de interes la distribución de probabilidad de una variable dado el resto de estas, usando la propiedad de la ecuación 2.5 se tiene que:

$$\begin{aligned} p(x_i, x_{j\neq i}) &= \frac{p(x_1, x_2, ..., x_D)}{\int_{x_i} p(x_1, x_2, ..., x_D) dx_i} \\ &= \frac{\prod_k p(x_k|pa_k)}{\int \prod_{x_k} p(x_k|pa_k) dx_i} \end{aligned} \tag{2.21}$$

Se aprecia que los factores de x_k que no involucran a ninguna variable x_i, pueden ser sacados de la integral y eventualmente se cancelarán con el producto de arriba.

De esta manera, los únicos factores que involucran a la variable x_i serán las que cumplan con las siguientes condiciones:

- Sean padres de x_i,
- Sean hijos de x_i
- Aquellas variables que compartan algún hijo con x_i.

Este conjunto de variables son conocidas como la "Cobija de Markov" o la "Frontera de Markov". Es decir, si se requiere encontrar la distribución condicional de una variable dado el resto de las variables existentes en un modelo de gráficas, esta distribución dependerá únicamente de las variables comprendidas en la Cobija de Markov.

En otras palabras, el factor solo dependerá de la variable, sus padres, sus hijos y los otros padres de todos sus hijos.

Dicho resultado suele ser muy útil cuando se requiere encontrar relaciones de independencia condicional en gráficas de este tipo, ya que permite a base de simple inspección establecer de forma más rápida la distribución requerida.

Un ejemplo de la Cobija de Markov se muestra en la figura 2.10.

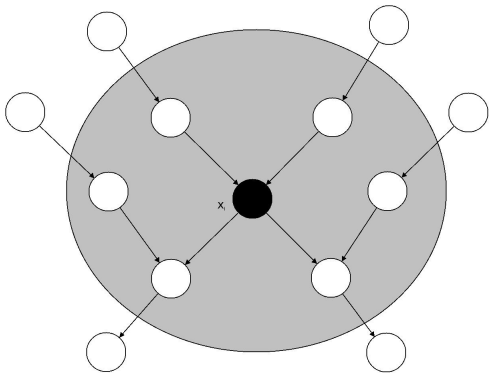


Figura 2.10: Un ejemplo de la Cobija de Markov para la variable x_i.

En resumen, a lo largo de este capítulo se ha dado una breve introducción y definición de modelos de gráficas probabilísticas. Se explicó la utilidad de estos en el área de la probabilidad y la estadística, además de presentar algunos ejemplos de cómo estas gráficas se pueden utilizar para mejorar la comprensión de un proceso que involucra múltiples variables aletorias. Posteriormente, se expone la definición de independencia condicional, usada en la simplificación de distribuciones conjuntas de probabilidad; así como, el concepto de "separación D" que resume las condiciones de independencia condicional. Finalmente, se define el concepto de "Cobija de Markov", el cual es un resultado importante que utiliza la independencia condicional para reducir la complejidad de expresiones condicionales de probabilidad.

Todos estos conceptos son necesarios para la comprensión del contenido de los capítulos 4 y 5, ya que a lo largo de estos se hace uso de conceptos y técnicas de modelos de gráficas probabilísticas. La idea de usar estas herramientas es tratar de simplicar las ecuaciones que representan a las distribuciones de probabilidad y encontrar expresiones que sean tratables para extraer la información deseada de los datos de manera eficiente.

Capítulo 3

Markov chain Monte Carlo (MCMC)

Por lo general, en computación existen una serie de métodos conocidos coloquialmente como "métodos numéricos", cuya función es estimar de forma iterativa valores de interés. En estadística una familia muy común para estimación de parámetros de lo modelos probabilísticos, es conocida con el nombre de métodos variacionales.

El problema del que adolecen los métodos variacionales y en general todo tipo de métodos numéricos, es la elección de los parámetros iniciales. Además, como se ha mencionado anteriormente, estos métodos proporcionan como salida una aproximación a los valores reales de los parámetros del modelo, y típicamente la precisión de la estimación está íntimamente relacionada con la elección de estos valores de inicio.

Debido a que la elección de estos puede ser un problema inclusive más complejo que la propia estimación de los parámetros del modelo, se han propuesto una serie de alternativas que gracias a su construcción carecen de este tipo de problemas.

Entre algunas de las propuestas para abordar esta problemática, se encuentra el realizar un muestreo de la distribución a priori. Los métodos Monte Carlo son precisamente una familia de métodos que usa esta estrategia para realizar la estimación de los parámetros.

Desde que los modelos gráficos del LDA y el de Autores y Tópicos requieren de la estimación de los hiperparámetros, los métodos Markov chain Monte Carlo (MCMC) permiten encontrarlos de forma eficiente, sin tener que lidiar con los problemas de elección de los valores iniciales y con una mejor precisión.

En el capítulo anterior se ha explicado como el manejo de los modelos de gráficas permite una manipulación más sencilla y ágil de las expresiones matemáticas que representan dichas distribuciones de probabilidad. Sin embargo, en la práctica, la inferencia bayesiana en muchos modelos de gráficas de interés es

intratable.

De acuerdo a lo mencionado en Gilks y Spiegelhalter (1995), el problema en general se encuentra en la necesidad de calcular integrales que en ocasiones son muy complejas de obtener de forma analítica. Para lidiar con este problema en muchos casos se recurre a aproximaciones numéricas, las cuales en algunos casos carecen de precisión, además de que en altas dimensiones se vuelven tan complejas que prácticamente son imposibles de resolver. Una forma de evitar estas situaciones es utilizar los métodos de muestreo.

Estos métodos tienen sus inicios en el campo de la física y su primera aparición se dio en Metropolis y Ulam (1949), pero hasta la década de los ochenta fue cuando comenzaron a ser ampliamente utilizados en el ramo de la estadística.

En general, estos métodos intentan encontrar la esperanza de una cierta función $f(\mathbf{z})$ únicamente muestreando de la distribución de probabilidad $p(\mathbf{z})$. Esto se puede calcular para variables continuas o discretas de la misma forma, simpleme cambiando las integrales por las sumatorias, dicha esperanza se puede expresar como:

$$\mathbb{E}(f) = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} \tag{3.1}$$

La idea principal es aproximar esta integral mediante el uso de la media aritmética, usando como datos conjunto de muestras independientes $\mathbf{z}^l$, es decir:

$$\hat{f} \approx \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^l) \tag{3.2}$$

Debido a la ley de los grandes números y a que $\mathbf{z}^l$ es muestreada de $p(\mathbf{z})$, entonces se puede garantizar que con un número suficientemente grande de L se cumple que $\mathbb{E}(f) = \mathbb{E}(\hat{f})$. Según Bishop (2007) se dice que en la práctica el número de muestras (valor de L) puede ser suficiente con diez o veinte.

De igual manera que con la media, se puede encontrar que la varianza de $f(\mathbf{z})$ está dada por:

$$Var(\hat{f}) = \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right] \tag{3.3}$$

En realidad el objetivo de los métodos MCMC es escoger dicha distribución de probabilidad $p(\mathbf{z})$ lo más simple y sencilla posible, tal que permita obtener de forma iterativa en cada paso una muestra candidata $\mathbf{z}^*$. Posteriormente, con un criterio definido con anterioridad, se debe tomar la decisión de si $\mathbf{z}^*$ debe será aceptada o bien rechazada.

De acuerdo con lo antes mencionado, los métodos MCMC constan básicamente de dos etapas complementarias:

- Usar cadenas de Markov (Markov chain) para obtener muestras $\mathbf{z}^t$ de la distribución de probabilidad.
- Después se realiza una aproximación del valor esperado de $f(\mathbf{z})$ mediante una integración al estilo Monte Carlo.

3.1. Cadenas de Markov

Las cadenas de Markov son modelos de gráficas, que representan un sistema que varía su estado a lo largo del tiempo. La principal característica de una cadena de Markov es que la probabilidad del obtener un valor en el siguiente estado, está influenciada por el valor obtenido en el estado presente.

El matemático ruso Andrei Andreevitch Markov (1856-1922) fue el primero en introducir el concepto de cadena de Markov, y tiene este nombre por la propiedad de encadenar los eventos, recordando el estado actual para poder influir en el estado futuro. Es decir, se cumple la ecuación 3.4.

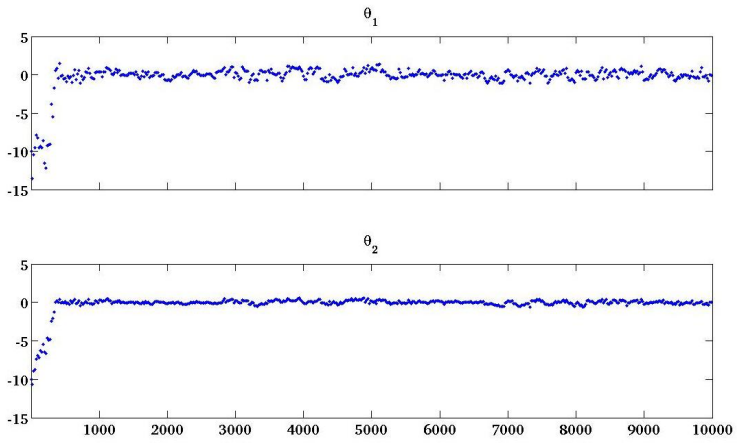
$$P(\mathbf{z}^{t+1}|\ \mathbf{z}^t, \mathbf{z}^{t-1}, ..., \mathbf{z}^1, \mathbf{z}^0) = P(\mathbf{z}^{t+1}|\mathbf{z}^t) \tag{3.4}$$

Cuando se obtiene una muestra $\mathbf{z}^1, \mathbf{z}^2 ..., \mathbf{z}^L$ se dice que la distribución condicional $P(\mathbf{z}^{t+1}|\mathbf{z}^t)$ es el "Kernel de Transición" de la cadena, debido a que dicha distribución de probabilidad define el comportamiento de la cadena.

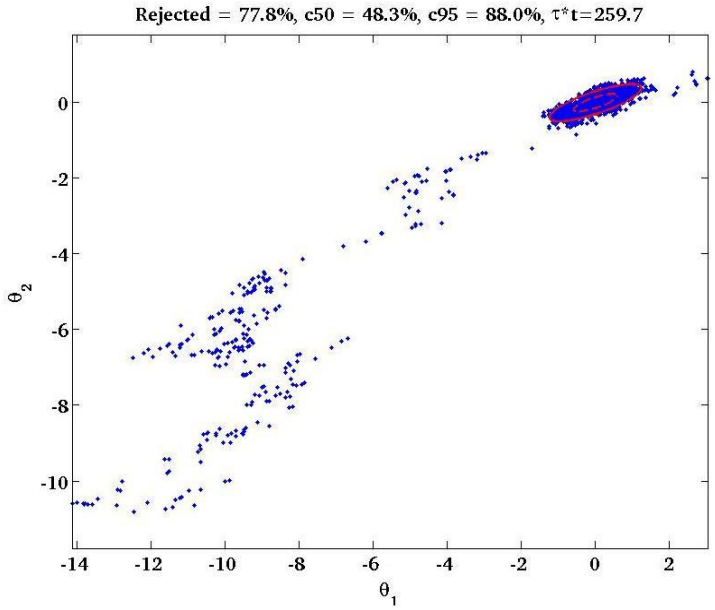
Gracias a la ecuación 3.4 se puede verificar que la cadena de Markov tiene la propiedad de que eventualmente, cuando $t \to \infty$ el resultado del estado futuro $\mathbf{z}^{t+1}$ ya no depende del historial de la cadena. En la práctica no se puede hacer $t = \infty$ sino solo encontrar un valor m suficientemente grande como para ocasionar, que a partir de ese momento la cadena de Markov olvide el estado inicial. Cuando esto ocurre se suele decir que la cadena de Markov ha convergido a su distribución estacionaria $\phi(\cdot)$. En el momento en que un conjunto de datos tiene la característica de "olvidar" el estado inicial cuando un proceso corre por mucho tiempo, se dice que tiene la propiedad de ergodicidad.

Cuando se ha convergido a la distribución estacionaria $\phi(\cdot)$ ocurre que el conjunto de muestras $\mathbf{z}^t;\ t = m+1, m+2...n$ serán dependientes y provendrán de la distribución estacionaria $\phi(\cdot)$. Se debe notar entonces que $m - n = L$, es decir, las muestras usadas en el cálculo del valor esperado deben ser tomadas después de la convergencia hacia la distribución estacionaria, para con estos datos proceder a realizar la integración Monte Carlo.

En las figuras 3.1a y 3.1b se aprecia la forma en la que se realiza el muestreo de una gaussiana multivariada con 2 parámetros θ_1 y θ_2 ambas iguales a 0. Note la concentración de puntos del lado superior derecho en donde se encuentra los valores de los parámetros adecuados para las gaussianas. Dos elipses de color rojo han sido dibujadas conteniendo en 95 y 50 % de las muestras tomadas a partir de la iteración 1000 usando 10000 iteraciones.



(a) Evolución de los parámetros



(b) Evolución del muestreo

Figura 3.1: Evolución de dos medias en el proceso de muestreo de la cadena de Markov.

3.2. Integración Monte Carlo

Una vez obtenido el conjuntos de muestras $\{\mathbf{z}^t; t = m+1, m+2...n\}$ se puede proceder a calcular la esperanza requerida de acuerdo con la ecuación 3.2. Ya que se muestrea de la distribución de probabilidad a priori, entonces es sencillo obtener tantas muestras como sean necesarias, y usando la ley de los grandes números, se puede manipular la presición de la aproximación. En la práctica según lo explicado en Bishop (2007) se dice que son suficientes tomar 10 o hasta 20 muestras.

3.3. Muestreo de Gibbs

El muestreo de Gibbs es uno de los algoritmos de MCMC más usados y es considerado como un caso especial del algoritmo Metropolis-Hasting. Se recurre a este algoritmo cuando se desconoce la distribución de probabilidad conjunta pero es posible estimar la distribución de probabilidad condicional de cada variable.

Se requiere muestrear de la distribución de probabilidad $p(\mathbf{z}) = p(z_1, z_2...z_M)$, por lo que en cada paso del muestreo de Gibbs se reemplaza el valor de cada variable por uno nuevo obtenido a través de un mestreo sobre la distribución a priori. Dicha distribución para el caso particular del muestreo de Gibbs, tiene la característica de estar condicionada sobre las restantes variables latentes.

Es decir, para cada variable z_i se muestrea de la distribución de probabilidad $p(z_i|z_{-i})$ donde $z_{-i} = z_1, z_2...z_M$ excluyendo z_i.

Por ejemplo, si se tiene una distribución $p(z_1, z_2, z_3)$ de tres variables y en un determinado paso τ se requiere muestrear el valor de $z_1^{\tau+1}$. Entonces se hará mediante la distribución de probabilidad condicional $p(z_1|z_2^\tau, z_3^\tau)$. Luego se obtiene el valor de $z_2^{\tau+1}$ de $p(z_2|z_1^{\tau+1}, z_3^\tau)$ y finalmente se obtendrá el valor de $z_3^{\tau+1}$ de $p(z_3|z_1^{\tau+1}, z_2^{\tau+1})$. Este procedimiento se repite cíclicamente hasta verificar la convergencia de la cadena hacia la distribución estacionaria.

Algoritmo 3.3.1 Algoritmo del muestreo de Gibbs.

0: Inicializamos $\{z_i; i = 1, 2...M\}$
para $\tau = 1$ to T **hacer**
  muestrear $z_1^{\tau+1} \sim p(z_1|z_2^\tau, z_3^\tau..., z_{M-1}^\tau, z_M^\tau)$
  muestrear $z_2^{\tau+1} \sim p(z_2|z_1^{\tau+1}, z_3^\tau..., z_{M-1}^\tau, z_M^\tau)$
  $\vdots$
  muestrear $z_{M-1}^{\tau+1} \sim p(z_{M-1}|z_2^\tau, z_3^\tau..., z_M^\tau)$
  muestrear $z_M^{\tau+1} \sim p(z|z_2^\tau, z_3^\tau..., z_{M-1}^\tau)$
fin para

La cadena de Markov debe de garantizar la propiedad de ergodicidad para asegurar la convergencia. Es posible demostrar que esta condición se cumple cuando en el conjunto de distribuciones condicionales usadas en el muestreo de Gibbs, son todas diferentes de cero. De no ser así para aplicar el método del muestreo Gibbs se requiere probar de forma explicita la ergodicidad.

La diferencia entre el muestro de Gibbs y el algoritmo Metropolis-Hasting, radica esencialmente en que Metropolis-Hasting se encarga de muestrear de la distribución de probabilidad condicional y luego decidir (como en la mayoría de los métodos MCMC) si la muestra candidata será aceptada.

Para hacer esto Metropolis-Hasting realiza el cálculo de la probabilidad de aceptación como se muestra en la ecuación 3.5.

$$A_k\left(z^*, z^\tau\right) = min\left(1, \frac{\tilde{p}(z^*)q_k(z^\tau|z^*)}{\tilde{p}(z^\tau)q_k(z^*|z^\tau)}\right) \tag{3.5}$$

Donde k representa una etiqueta de transición de un estado a otro y q_k es la distribución condicional en el estado K.

Después se genera un número aleatorio $u\,\epsilon\,[0,1]$ y si $u > A_k$ la muestra candidata es rechazada de lo contrario es agregada.

Para el caso particular del método del muestro de Gibbs, esta probabilidad de aceptación se simplifica como se muestra a continuación:

$$\begin{aligned} A_k\left(z^*, z^\tau\right) &= min\left(1, \frac{\tilde{p}(z^*)q_k(z^\tau|z^*)}{\tilde{p}(z^\tau)q_k(z^*|z^\tau)}\right) \\ &= min\left(1, \frac{p(z_k^*|z_{-k}^*)p(z_{-k}^*)p(z_k|z_{-k}^*)}{p(z_k|z_{-k})p(z_{-k})p(z_k|z_{-k})}\right) \\ &= min\left(1, 1\right) \\ &= 1 \end{aligned} \tag{3.6}$$

Es decir, en cada paso del muestreo de Gibbs a diferencia del Metropolis-Hasting la muestra candidata z^* siempre es aceptada.

Finalmente, es importante mencionar que la eficiencia del algoritmo del muestreo de Gibbs depende en gran parte de la complejidad de la distribución de probabilidad condicional $P(z_i|z_{-i})$. Cuando dicha distribución de probabilidad está especificada por un modelo gráfico entonces las variables involucradas son aquellas que estén comprendidas en la Cobija de Markov .

3.4. Dificultades para hacer inferencia usando el MCMC

A pesar de que el uso de métodos Monte Carlo, proporciona un marco de trabajo para la aproximación de distribuciones de probabilidad, es posible que se incurra en algunos errores en el cálculo de dicha aproximación, algunas de las causas que producen errores son:

- Se puede obtener un modelo inapropiado: El modelo podría no ajustar a los datos o bien podría no ser realista.
- Errores de cálculo o programación: Las diferentes implementaciones de los algoritmos así como, algunos problemas de errores numéricos pueden ocasionar que la distribución estacionaria calculada no sea la adecuada.
- Convergencia lenta: Debido a la influencia de los puntos de inicio establecidos para los algoritmos, es posible que la simulación permanezca por muchas iteraciones en alguna región de espacio de búsqueda. Si el criterio de paro está basado en el número de iteraciones, el algoritmo dará como resultado un modelo que aún no ha convergido a la distribución estacionaria y por lo tanto sugerirá un modelo erróneo.

En general, los tres errores suelen aparecer de forma constante en cualquier método de estimación estadístico, sin embargo los primeros dos, suelen ser muy comunes en los métodos MCMC, debido a la complejidad de la programación para las cadenas de Markov. Por otro lado la convergencia lenta suele ser un problema que ocurre en métodos tales como el algoritmo *Expectation Maximization (EM)*, que de igual manera ocasionan errores al aceptar modelos cuya convergencia no ha sido garantizada y únicamente convergen a un máximo local.

Alguna de las sugerencias para lidiar con los problemas mencionados, es monitorear estadísticos que resuman el estado de la cadena y también el replicar los experimentos con diferentes puntos iniciales, con la esperanza de encontrar convergencia hacia la misma solución o bien ir descubriendo soluciones múltiples si las hay. Sin embargo, a diferencia de los métodos deterministas, los métodos MCMC complican la implementación de estas estrategias, ya que al ser de naturaleza estocástica los estadísticos de resumen de estado no necesariamente son monotónicamente crecientes o decrecientes. Además es mucho más difícil hacer comparaciones entre los resultados, ya que estos son distribuciones de probabilidad y no puntos en $\mathbb{R}^n$.

3.4.1. Monitoreo de convergencia y problemas ocasionados por la convergencia lenta

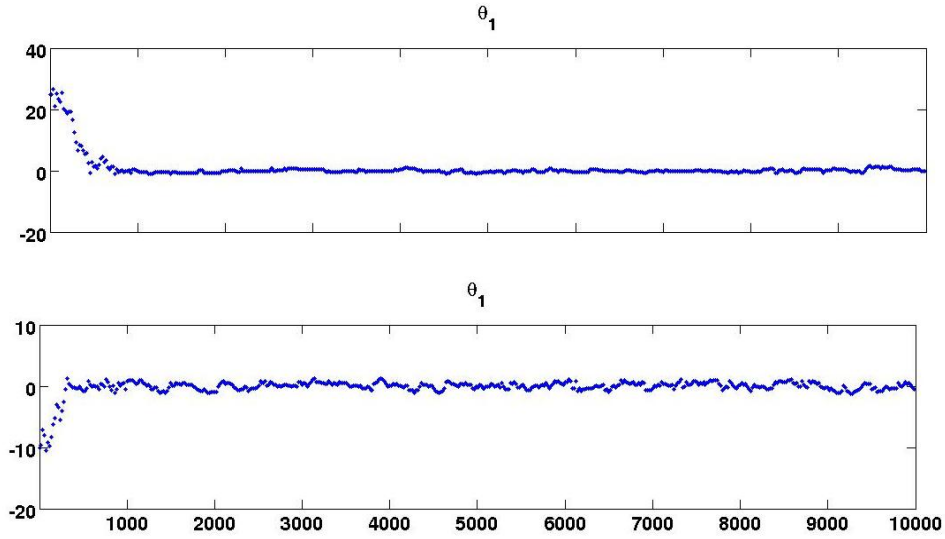
A pesar de garantizar la ergodicidad en los algoritmos de MCMC, no es posible garantizar un recorrido completo a través de todo el espacio de búsqueda en un número pequeño de iteraciones. Debido a esto, es posible que ocurran situaciones en la cual un algoritmo MCMC, se encuentre estacionado en alguna región del espacio de búsqueda, dando la impresión de que la cadena ha convergido a la distribución estacionaria. Puede suceder, sin embargo, que no se haya encontrado la verdadera distribución estacionaria, debido a que ni siquiera se ha visitado la región del espacio de búsqueda en la cual se encuentra.

Por este motivo es arriesgado tomar decisiones acerca de la convergencia de un algoritmo MCMC basado únicamente en una sola cadena, ya que puede parecer que ha convergido perfectamente, cuando en realidad el algoritmo únicamente se ha quedado atascado en una lenta convergencia de una región no adecuada del espacio de búsqueda.

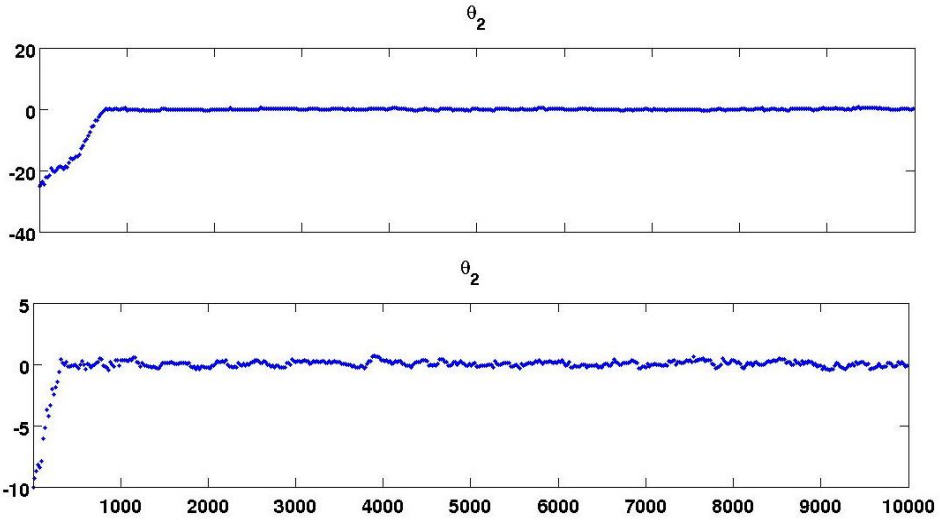
Es importante mencionar que los criterios de convergencia deben cumplirse para cada uno de los parámetros a estimar. Por lo tanto, el proceso de muestreo no debe ser interrumpido mientras aún exista algún parámetro en el que no se garantise que la cadena ya ha convergido a la distribución estacionaria.

Por esta razón, siempre es recomendable que se corran varias cadenas inicializadas en diferentes puntos, permitiendo hacer comparaciones entre las evoluciones en el tiempo. Es deseable que además los puntos iniciales estén adecuadamente distribuidos en el espacio de búsqueda, tratando de explorar en todas las regiones donde la solución puede estar. Esto es posible la mayoría de las veces cuando se trabaja con parámetros discretos o cuyos valores están acotados. Sin embargo, en ocasiones el proceso de correr múltiples cadenas resulta computacionalmente muy costoso (Gelman y Rubin (1992) y Geyer (1992)).

En la figura 3.2 se puede observar la evolución de los parámetros θ_1 y θ_2 del ejemplo perteneciente a la estimación de medias para una gaussina multivariada de la figura 3.1, en 2 cadenas diferentes y con puntos de inicialización distintos. La figura 3.2a presenta la evolución de la cadena con puntos de inicialización 25 y -10 respectivamente. La figura 3.2b ejemplifica la evolución del parámetro θ_2 con valores de inicialización -25 y -10. Se debe observar que en todas las gráficas se comienzan a percibir patrones a partir de la iteración 1000 en adelante, demostrando que las cadenas han convergido a la distribución estacionaria a partir de dicha iteración.



(a) Evolución θ_1



(b) Evolución θ_2

Figura 3.2: Evolución de los parámetros θ_1 y θ_2 inicializados con diferentes valores.

Es recomendable monitorear el comportamiento de la cadena (aún cuando aparentemente haya convergido) por un período largo de tiempo, ya que en general, es posible que la cadena no haya convergido aún, sino que simplemente se atoró en una región equivocada. Un ejemplo de esta situación se presenta en la figura 3.3, donde se aprecia una aparente convergencia del parámetro θ_1 cerca de la iteración 1000, pero cerca de la iteración 8,600, el algoritmo logra salir del área de muestreo y comienza a explorar en otras secciones del espacio de búsqueda.

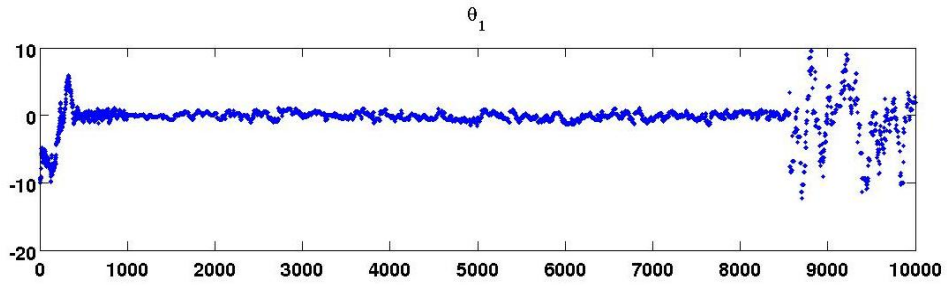


Figura 3.3: Una aparente convergencia del parámetro.

Algunas estrategias útiles para diagnosticar la convergencia de un algoritmo Monte Carlo se presentan a continuación.

3.4.2. Estudio de la autocorrelación

La intención del proceso de muestreo a través de una cadena de Markov, es eventualmente obtener muestras de la distribución estacionaria. Esto de acuerdo al procedimiento de los algoritmos MCMC, sugiere que eventualmente la cadena de Markov "olvidará" su valor inicial y comenzará a muestrear de la distribución estacionaria.

El estudio de la autocorrelación está basado precisamente en el hecho de que las muestras obtenidas con la cadena de Markov, mantendrán un alta correlación al principio del muestreo, y esta disminuirá con el aumento de las iteraciones. Es decir, si se estudia la correlación de las muestras de cada parámetro, obtenidas desde el tiempo t hasta el tiempo $t+n$ con una n suficientemente grande (típicamente $n = 50$), es posible analizar la convergencia de la cadena al observar una gráfica de la evolución de la autocorrelación monotónicamente decreciente y con valores de autocorrelación bajos preferentemente.

Una gráfica de autocorrelación monotónicamente decreciente, da indicios de que el proceso de muestreo ha convergido a una distribución estacionaria. Por otro lado, la convergencia puede ser engañosa como se ha discutido anterioemente. Un mayor indicio de que se ha convergido a la distribución estacionaria adecuada, se presenta al obtener valores de correlación bajos por largos períodos de tiempo.

Nuevamente el concepto de períodos de muestreo largos es ambiguo. Ya que dependerá totalmente del comportamiento de la cadena y de la distribución estacionaria.

La figura 3.4 muestra una distribución que aparentemente ha convergido, pero un valor aún por arriba de 0.75 indica que la cadena aún debe continuar siendo muestreada para converger a la distribución estacionaria.

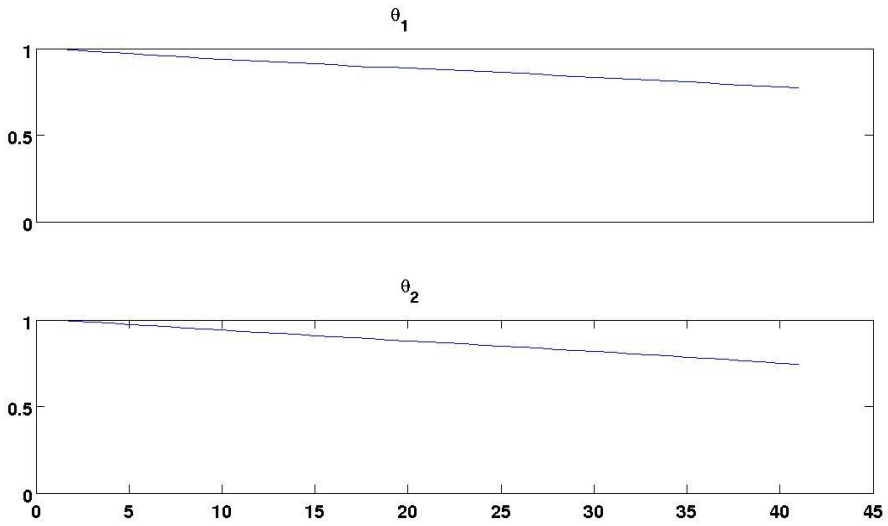


Figura 3.4: Gráfica de la autocorrelación para los parámetros θ_1 y θ_2 con una correlación aún elevada.

La figura 3.5 presenta una gráfica de la autocorrelación que sugiere una distribución estacionaria y con valores cercanos a 0.5, indicando una mejor convergencia.

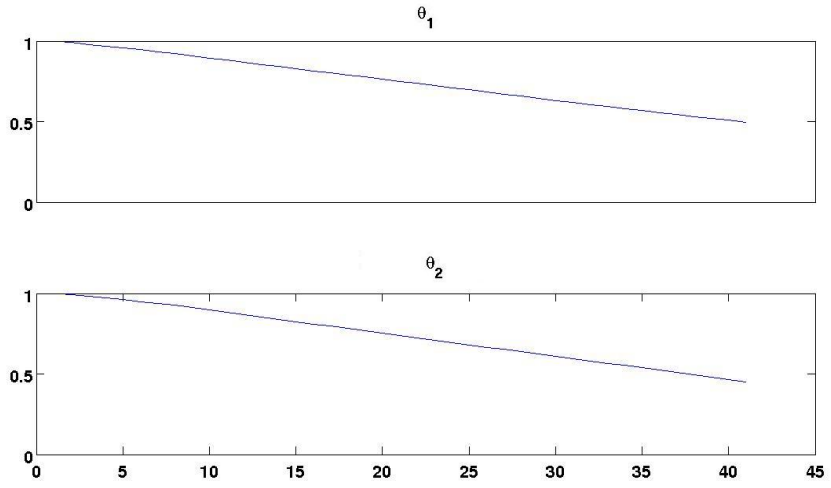


Figura 3.5: Gráfica de la autocorrelación para los parámetros θ_1 y θ_2 con una correlación moderada.

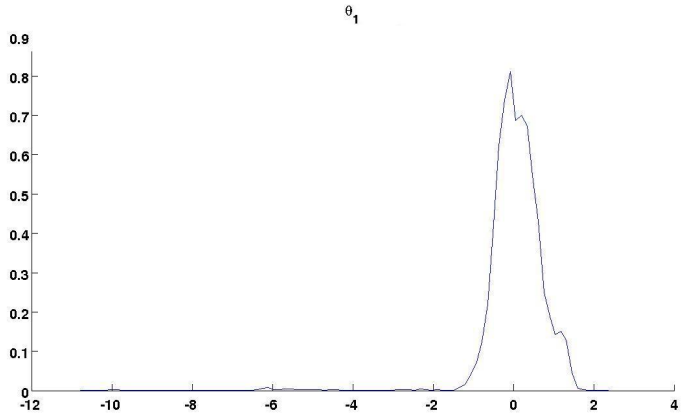
3.4.3. Gráficas de Kernel

Otra técnica de visualización bastante común para el diagnóstico de la convergencia se basa en la estimación de kernels para los parámetros. En general, el

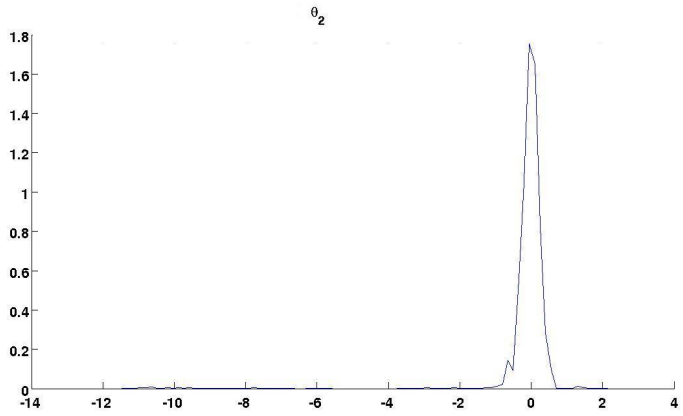
kernel para un parámetro no debe tener más de una moda, el obtener gráficas multimodales es un mal síntoma de una no convergencia. Por otro lado, el obtener gráficas de kernels monomodales pero no bien definidas, es un indicador de que el muestreo ya se encuentra estacionado en una sección del espacio de búsqueda, pero aún no ha convergido del todo.

Considerando el ejemplo de la gaussiana multivariada expuesto en la figura 3.1, en la 3.6 se presenta una gráfica de estimación de kernels para los parámetros θ_1 y θ_2. tomadas en tempranas iteraciones del proceso de muestreo.

En las figuras 3.6a y 3.6b se aprecian diversas deformaciones en la gráfica de kernels. También demuestran que el muestreo ya se ha estacionado en una sección del espacio de búsqueda, pero aún no ha convergido del todo.



(a) Estimación del Kernel para θ_1

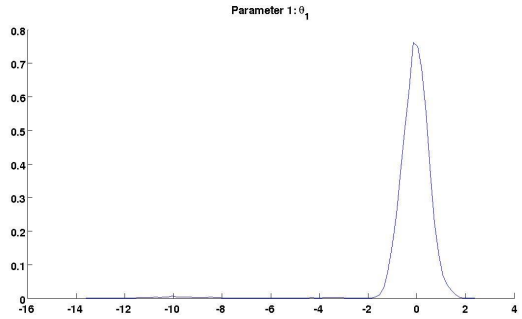


(b) Estimación del Kernel para θ_2

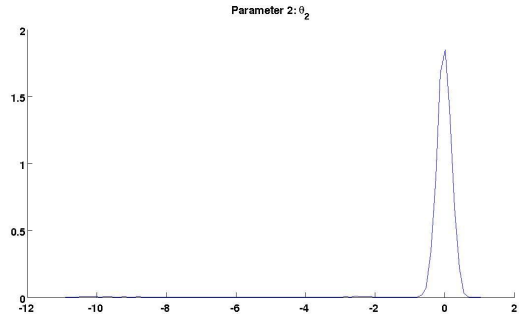
Figura 3.6: Estimación de kernels que aún no demuestran convergencia.

Por otro lado la figura 3.7 fue tomada casi al final del proceso de muestreo

cuando ya se ha convergido a la distribución estacionaria. Ambas figuras muestran kernels gaussianos bien definidos por lo que demuestran convergencia adecuada.



(a) Estimación del Kernel para θ_1



(b) Estimación del Kernel para θ_2

Figura 3.7: Estimación de kernels que demuestran convergencia.

3.4.4. Prueba Z de Geweke

Este método realiza una prueba estadística Z sobre las medias de dos poblaciones generadas a partir de la primera y segunda mitad de las muestras que devuelve el algoritmo MCMC, según se explica en Geweke (1992).

La idea a trabajar se basa en que en el momento en que la cadena ha convergido, si se toman muestras de la primera y segunda parte de la cadena y se calculan las medias, entonces es posible compararlas de forma estadística.

Las poblaciones con las que se calculan las medias son tomadas aleatoriamente en una relación típicamente de un 10 % de las muestras de la primera mitad de la cadena y un 50 % de la segunda. Es claro que la primera mitad de la cadena está en desventaja con respecto a la segunda, ya que la primera parte contiene las muestras cuando la cadena aún no ha convergido. Por esta razón solo se toman un 10 % de la población mientras que la segunda parte de la cadena se forzó a tener convergencia al tomar el 50 % de la muestra.

Una vez tomadas las muestras se puede calcular un valor promedio para cada mitad de los parámetros y aplicar una prueba de hipótesis determinando si ambas medias son iguales estadísticamente.

De ser diferentes las medias, la cadena no ha convergido y se debe dejar correr por mayor tiempo. Si son iguales, se puede tomar como parámetro de "burn in" el número de iteración que divide a la cadena, puesto que cualquier muestra tomada en un iteración posterior a la mitad provendrá de la distribución estacionaria.

3.4.5. Método de Gelman-Rubin

Este enfoque es el considerado el mejor de todos y se dio a conocer inicialmete en Gelman y Rubin (1992). Consiste en monitorear la convergencia mediante el uso de cantidades escalares que son llamados indicadores, los cuales suelen ser de sumo interés cuando se trata de monitoreo de convergencia. Por ejemplo, en ocasiones es de interés realizar la estimación de varios parámetros de una distribución, mediante el uso de un indicador por cada parámetro, es posible realizar el monitoreo de convergencia de forma individual.

Para esto se nombra a cada uno de los indicadores como ψ. Y suponga que para el procedimiento se han escogido correr m cadenas cada una de longitud n, entonces se obtiene un matriz, donde cada elemento $\psi_{i,j}$ es un indicador de interés tal que $i = 1, 2, ..., n$ y $j = 1, 2, ...m$. Posteriormente, se definen las siguientes cantidades escalares:

$$B = \frac{n}{m-1}\sum_{i=1}^{m}\left(\overline{\psi_{i.}} - \overline{\psi_{..}}\right)^2, \tag{3.7}$$

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2, \;\; donde \;\; s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\psi_{ij} - \overline{\psi_{i.}}\right)^2 \tag{3.8}$$

La ecuación 3.7 representa la varianza entre las secuencias, mientras que la ecuación 3.8 indica la varianza intra secuencia. Ahora se calcula un estimador de la varianza que funcione para medir la varianza en general intra y entre secuencias como:

$$\widehat{var}\left(\psi\right) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{3.9}$$

La varianza general de la ecuación 3.9 puede ser manipulada mediante el uso de los puntos iniciales de las secuencias. Cuando las secuencias se inicializan con

puntos iniciales sobre dispersos, entonces el estadístico de la ecuación 3.9 aumenta, siendo el valor de este sobreestimado. Por otro lado, el valor de W mientras no se haya convergido a la distribución estacionaria se dice que subestima la varianza de ψ.

Si se define la razón siguiente:

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{var}\,(\psi)}{W}} \tag{3.10}$$

Se puede encontrar un porcentaje en la que la varianza sobreestimada supera el valor de la varianza subestimada. En teoría cuando se deja correr por mucho tiempo ($n \to \infty$) se debe obtener un valor de $\sqrt{\widehat{R}} = 1$. En la práctica es casi imposible obtener tal grado de convergencia, sin embargo, los valores por debajo de 1.2 son admisibles como prueba de una adecuada convergencia.

Es necesario aclarar que para poder obtener una estimación aceptable de los parámetros, se debe correr una cantidad adecuada de secuencias, siendo el valor de $m = 10$ típicamente en la práctica. Pero un un número mayor para m, podría ayudar a calcular de forma más eficiente el valor adecuado de las varianzas, aunque el costo computacional aumenta.

A modo de repaso, en este capítulo se ha presentado una introducción a los métodos de muestreo MCMC, en especial el muestreo de Gibbs. Se ha explicado su funcionamiento y porque se constituyen como una alternativa a los métodos de estimación bayesiana comúnmente usados. De igual manera, se comentó acerca de los inconvenientes que esta técnica suele tener, en especial los problemas relacionados con el monitoreo de la convergencia de la cadena de Markov. Sobre este último tema, se profundizó presentando una serie de estrategias para verificar la convergencia del algoritmo.

En capítulos posteriores, se emplea el muestreo de Gibbs para realizar la asignación de las variables latentes y la estimación de los parámetros para los modelos LDA y de Autores y Tópicos.

Capítulo 4

Latent Dirichlet Allocation (LDA)

En este capítulo se describe el modelo probabilístico generativo "Latent Dirichlet Allocation" (LDA). Este permite a un conjunto de observaciones ser explicadas a través del uso grupo de variables latentes.

El modelo LDA hace la asunción de una bolsa de palabras, es decir, el orden de aparición de estás es irrelevante. Mediante el uso de los llamados tópicos, el LDA intenta encontrar la forma en la que las palabras fueron apareciendo, dando lugar a las observaciones.

Un tópico, desde el punto de vista estadístico, es una distribución de probabilidad sobre las palabras, típicamente usando una multinomial. Los modelos de tópicos, como su nombre lo indica, tratan de modelar la aparición de las palabras asignando la probabilidad de aparición de cada palabra para un tópico específico. Para poder reflejar de forma adecuada el proceso en el cual se generó la colección, se recurre a definir un conjunto de tópicos, tantos como sean necesarios. Además, también se modela la forma de los documentos, por medio de una mezcla de tópicos, es decir, un documento se conforma por un porcentaje de un conjunto predefinido de tópicos.

El modelo LDA fue dado a conocer inicialmente en Blei *et al.* (2003) y en sus comienzos fue usado para el análisis de grandes colecciones de documentos de texto. Sin embargo, hoy día se encuentran implementaciones para diversas áreas, es el caso de Elango y Jayaraman (2005) para agrupamiento de imágenes, aprendizaje no supervisado de clases de objetos como en Endres *et al.* (2009), categorización de documentos expuesto en Yang (1997), entre otras aplicaciones.

En el contexto de procesamientos de textos, la naturaleza continuamente cambiante de algunas colecciones de documentos imposibilitan la clasificación manual. Por otro lado, técnicas de aprendizaje no supervisado, permiten una mayor eficiencia, al descubrir la estructura de la información de forma automatizada y sin necesidad de intervención humana.

Una técnica de aprendizaje no supervisado como lo es el *Análisis de Semán-*

tica Latente (LSA) por sus siglas en inglés, está descrito por primera vez en Deerwester *et al.* (1990). Emplean factorizaciones de la matriz de ocurrencias de las palabras sobre los documentos. Intenta capturar la información relevante acerca de la estructura de los datos, descomponiendo dicha matriz y luego resumiendo la información mediante un análisis de componentes f principales.

El *Análisis Probabilístico de la Semántica Latente (PLSA)*, por otro lado, fue dado a conocer en Thomas (1999), y representa una mejora del LSA. Sugiere un modelo generativo basado en variables aleatorias en lugar de considerar únicamente interpretaciones geométricas de las características de la información. Supone que cada palabra proviene de un tópico seleccionado y que cada uno de los documento en la colección está formado por una mezcla de varios tópicos.

El modelo LDA extiende esta idea, considerando que la aparición de los tópicos está regida por una distribución de de probabilidad multinomial, que a su vez tiene un prior Dirichlet. El LDA proporciona un marco confiable para realizar un proceso de aprendizaje no supervisado, que permite analizar la estructura de la información observada.

4.1. Modelo de gráficas probabilísticas y proceso generativo del LDA

El modelo de gráficas probabilísticas que describe el proceso generativo para el LDA se presenta en la figura 4.1.

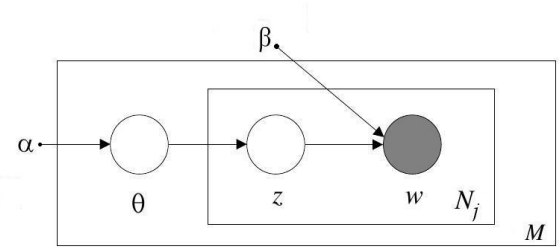


Figura 4.1: Modelo generativo probabilístico LDA.

Consta de dos fases que generan un documento:

- En la primera fase se selecciona el parámetro θ proveniente de una distribución Dirichlet con hiperparámetro α. θ representa la distribución multinomial de los tópicos para este documento. Este proceso de selección se realiza

para cada uno de los documentos en la colección, por lo que de acuerdo a lo especificado en la figura 4.1 este proceso se realiza M veces.

- Una vez seleccionada la distribución multinomial entonces se muestrea de ella z. Con este valor, ahora se muestrea la palabra seleccionando la distribución de probabilidad del parámetro β, tomándola de la i-ésima fila de esta matriz. Es decir, β es una matriz con tantas filas como tópicos existan en la colección y tantas columnas como palabras en el diccionario. Además, cada fila de la matriz representa una distribución multinomial de las palabras sobre los tópicos. Este procedimiento se repite para cada una de las N_j palabras en el documento j.

El algoritmo para este proceso es el siguiente:

Algoritmo 4.1.1 LDA.

para todo Documento en la colección **hacer**
  Muestrear $\theta_j \sim Dir(\alpha)$
  para i=1 hasta N_j **hacer**
    Muestrear $z_{j,i} \sim Mult(\theta)$
    Muestrear $w_{j,i}$ de $P(w_{j,i}|z_{j,i}, \beta)$ una distribución de probabilidad Multinomial condicionada en $z_{j.i}$
  fin para
fin para

La finalidad de este modelo al igual que el PLSA es estimar los valores de los parámetros θ y β que con mayor probabilidad generaron los datos observados. Para lograr este objetivo se han propuesto múltiples métodos, entre los que sobresalen la estimación mediante el algoritmo EM.

4.2. LDA suavizado

Es un hecho conocido que en colecciones de documentos existen muchas palabras que aparecen únicamente en un determinado documento y no ocurren nuevamente en ningún otro.

Un problema que presenta el modelo LDA, está relacionado con este hecho. El procedimiento de estimación de parámetros utilizando inicialmente (conocido como el valor máximo de la verosimilitud) en un conjunto de entrenamiento, asignará la probabilidad de cero a palabras que no se encuentren en este conjunto, ocasionando que un nuevo documento con dicha palabra tenga una probabilidad de cero también en lugar de incrementarla. La solución propuesta en Blei *et al.*

(2003) es agregar una nueva variable aleatoria al modelo, conocida con el nombre de "variable de suavizado" cuya función es asignar a todas las palabras de la colección una pequeña probabilidad.

En el caso de la resolución de este modelo usando el método del muestreo de Gibbs explicado en la sección 3.3, es importante recordar que se requiere que las probabilidades de transición de los estados en la cadena de Markov sean todas diferentes de cero. Posteriormente, se observará al desarrollar las ecuaciones para la estimación de los parámetros del modelo, que gracias a estas variables de suavizado es posible garantizar la ergodicidad del modelo y poder obtener siempre una aproximación para los parámetros.

El modelo que se obtiene de esta variante del modelo LDA, es conocido como modelo LDA suavizado y se presenta en la figura 4.2.

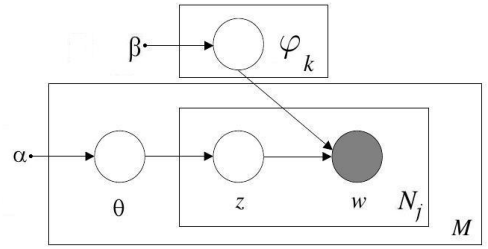


Figura 4.2: Modelo generativo probabilístico LDA suavizado.

Esta nueva variable φ es en realidad una matriz $K \times V$ donde K es el número de tópicos y V es el número de palabras en el vocabulario. Cada fila de la matriz φ representa una distribución de probabilidad multinomial sobre las palabras y se seleccionan con base en el valor obtenido al muestrear la variable aleatoria z. También a través del modelo de gráficas, se hace notar que cada una de las filas de la variable φ es independiente de las demás y se forma al muestrear de una distribución Dirichlet con parámetro β, esto es $\varphi_k \sim Dir\,(\beta)$ donde β es un vector de longitud V.

El proceso generativo de este modelo es el mismo al descrito en el algoritmo 4.1.1, salvo que en esta ocasión, la palabra no será obtenida directamente de una multinomial con parámetro β. Por el contrario, β en esta ocasión será el priori de la distribución multinomial φ_k.

Observando la figura 4.2 se puede deducir la verosimilitud de este modelo que se escribe como:

$$P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta) = P(\boldsymbol{\theta}|\alpha)P(\boldsymbol{z}|\boldsymbol{\theta})P(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\varphi})P(\varphi|\beta) \tag{4.1}$$

Donde las variables escritas en negritas representan la versión vectorial o inclusive matricial de cada una de las respectivas variables. Nótese que debido al

plate notation de la figura 4.2 es posible deducir propiedades de independencia condicional de las variables y θ_j y θ_i si $i \neq j$. Este es el mismo caso para cada una de las variables $\mathbf{z}$,$\mathbf{w}$ y $\boldsymbol{\varphi}$. Por último, se puede tomar el producto de cada una de las variables antes descritas sobre sus respectivos índices y volver a expresar la ecuación 4.1 de la siguiente forma:

$$\begin{aligned}
&P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta) \\
&= \left(\prod_{j=1}^{M} P(\theta_j|\alpha) \left[\prod_{i=1}^{N_j} P(z_{j,i}|\theta_j) P(w_{j,i}|\varphi_{z_{j,i})}\right]\right) \left(\prod_{k=1}^{K} P(\varphi_k|\beta)\right) \\
&= \left(\prod_{j=1}^{M} P(\theta_j|\alpha) \prod_{i=1}^{N_j} P(z_{j,i}|\theta_j)\right) \left(\prod_{j=1}^{M}\prod_{i=1}^{N_j} P(w_{j,i}|\varphi_{z_{j,i})})\right) \left(\prod_{k=1}^{K} P(\varphi_k|\beta)\right) \\
&= \left(\prod_{j=1}^{M} P(\theta_j|\alpha) \prod_{i=1}^{N_j} P(z_{j,i}|\theta_j)\right) \left(\prod_{j=1}^{M}\prod_{i=1}^{N_j} P(w_{j,i}|\varphi_{z_{j,i})}) \prod_{k=1}^{K} P(\varphi_k|\beta)\right)
\end{aligned} \tag{4.2}$$

Ahora se harán algunas precisiones acerca de la notación utilizada. Se define:

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_M\} \text{ y } \theta_j = \{\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,K}\}$$

Cada θ_j representa una distribución multinomial de probabilidad del documento j en los K tópicos. Es importante recordar que $\theta_j \sim Dir(\alpha)$.

Así mismo, se definirá $z_{j,i}$ que representa el tópico al que pertenece la i-ésima palabra en el j-ésimo documento. Cada $z_{j,i}$ es una variable indicadora en notación 1 de k, es decir, tiene la forma:

$$z_{j,i} = [z_{j,i,1} = 0, z_{j,i,2} = 0, \ldots, z_{j,i,k} = 1, \ldots, z_{j,i,K} = 0] \text{ para alguna } k.$$

Como se puede observar, solo el k-ésimo componente es 1 mientras que los demás son 0. En notación matemática significa que $\sum_{k=1}^{K} z_{j,i,k} = 1$. El propósito de esta variable es definir mediante el uso del componente qué tópico ha sido seleccionado para contribuir con la i-ésima palabra en el j-ésimo documento muestreando de la distribución multinomial con parámetro θ_j.

Análogamente se define $w_{j,i}$ como la i-ésima palabra en el j-ésimo documento.

De igual manera que lo hace $z_{j,i}$, la variable $w_{j,i}$ es una variable indicadora por lo que tiene la siguiente forma:

$$w_{j,i} = [w_{j,i,1} = 0, w_{j,i,2} = 0, \ldots, w_{j,i,v} = 1, \ldots, w_{j,i,V} = 0] \text{ para alguna } v.$$

Queda claro gracias al modelo de gráficas, que la selección de esta palabra depende del k-ésimo tópico que se haya seleccionado y de la k-ésima distribución de probabilidad multinomial φ_k obtenida al muestrear de la distribución Dirichlet con parámetro β. Debido a esta explicación la variable φ entonces es una matriz tal que:

$$\varphi = \begin{pmatrix} \varphi_{1,1} & \varphi_{1,2} & \varphi_{1,V} \\ \varphi_{2,1} & \varphi_{2,2} & \varphi_{2,V} \\ \vdots & \ddots & \vdots \\ \varphi_{k,1} & \varphi_{k,2} & \varphi_{k,V} \\ \vdots & \ddots & \vdots \\ \varphi_{K,1} & \varphi_{K,2} & \varphi_{K,V} \end{pmatrix}$$

La k-ésima fila de la variable φ ha sido marcada para ejemplificar el proceso que se sigue al seleccionar una determinada palabra cuando el k-ésimo tópico ha sido elegido.

Es importante recordar que:

$$P(\varphi_k|\beta) = \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\prod_{v=1}^{V}\varphi_{k,v}^{\beta_v-1} \tag{4.3}$$

al igual que:

$$P(\theta_j|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{j,k}^{\alpha_k-1} \tag{4.4}$$

Esto es debido a que $\varphi_k \sim Dir(\beta)$ y $\theta_j \sim Dir(\alpha)$.

Una vez aclarado todo esto se puede verificar que:

$$P(z_{j,i}|\theta_j) = \prod_{l=1}^{K}\theta_{j,k}^{z_{j,i,l}} = \theta_{j,k} \tag{4.5}$$

Ya que la variable aleatoria $z_{j,i,l} = 1$ y todas las demás $z_{j,i,l} = 0$ si $l \neq k$ salvo en el índice k. Esto anulará todas las variables $\theta_{j,k}$ con $l \neq k$ elevándolas a la 0, mientras que la única variable que permanecerá es la $\theta_{j,k}$ que será elevada a la 1.

La ecuación 4.5 puede ser reescrita nuevamente como se presenta a continuación:

$$P(z_{j,i}|\theta_j) = \prod_{l=1}^{K}\prod_{r=1}^{V}\theta_{j,k}^{z_{j,i,l}x_{j,i,r}} \tag{4.6}$$

El cambio del índice r no afecta al índice l, por lo tanto $x_{j,i,r}$ tomará el valor de 1 hasta que $r = v$. De igual manera, $z_{j,i,l} = 1$ si y solamente si $l = k$. El producto $z_{j,i,l}x_{j,i,r} = 1$ únicamente cuando $r = v$ y $l = k$, lo cual significa, que la potencia solo toma el valor de 1 si los valores de r y l son v y k respectivamente.

De forma similar se puede definir $P(w_{j,i}|\varphi_{z_{j,i}})$ como:

$$P(w_{j,i}|\varphi_{z_{j,i}}) = \prod_{l=1}^{K}\prod_{r=1}^{V}\varphi_{k,v}^{z_{j,i,l}x_{j,i,r}} \tag{4.7}$$

Sustituyendo las ecuaciones 4.3, 4.4, 4.6 y 4.7 en la ecuación 4.2 se obtiene:

$$\begin{aligned} &P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta) \\ &= \left(\prod_{j=1}^{M} P(\theta_j|\alpha)\prod_{i=1}^{N_j} P(z_{j,i}|\theta_j)\right)\left(\prod_{j=1}^{M}\prod_{i=1}^{N_j} P(w_{j,i}|\varphi_{z_{j,i}})\prod_{k=1}^{K} P(\varphi_k|\beta)\right) \end{aligned} \tag{4.8}$$

$$= \left(\prod_{j=1}^{M}\left[\frac{\Gamma\left(\sum_{l=1}^{K}\alpha_l\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_l\right)}\prod_{l=1}^{K}\theta_{j,l}^{\alpha_l-1}\right]\prod_{i=1}^{N_j}\left[\prod_{l=1}^{K}\prod_{r=1}^{V}\theta_{j,l}^{z_{j,i,l}x_{j,i,r}}\right]\right) \tag{4.9}$$

$$\left(\prod_{j=1}^{M}\prod_{i=1}^{N_j}\left[\prod_{l=1}^{K}\prod_{r=1}^{V}\varphi_{l,r}^{z_{j,i,l}x_{j,i,r}}\right]\prod_{l=1}^{K}\left[\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma\left(\beta_r\right)}\prod_{r=1}^{V}\varphi_{l,r}^{\beta_r-1}\right]\right) \tag{4.10}$$

Es notorio que en la ecuación 4.9 y 4.10 el índice i no aparece en ningún lado más que en el exponente. Simplificando esta expresión se pueden reescribir estas ecuaciones de la siguiente forma:

- Denota $n_{j,l,r} = \sum_{i=1}^{N_j} z_{j,i,l} x_{j,i,r}$
- Si se suma el término $n_{j,l,r}$ sobre alguna variable, como por ejemplo r, se denotará como $n_{j,l,\cdot} = \sum_{r=1}^{V} n_{j,l,r}$.

Llevando a cabo dichas simplificaciones y substituciones al tomar el producto de las ecuaciones 4.9 y 4.10 se expresa la ecuación 4.7 usando la notación antes mencionada como:

$$\begin{aligned} & P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \beta) \\ &= \left(\prod_{j=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{l=1}^{K} \theta_{j,k}^{n_{j,l,\cdot}+\alpha_k-1} \right) \left(\prod_{l=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_v-1} \right) \end{aligned} \tag{4.11}$$

Ahora bien, dado que se puede suponer que $\alpha = [\alpha_0, \alpha_0 ..., \alpha_0]$ y $\beta = [\beta_0, \beta_0, ..., \beta_0]$ entonces la ecuación 4.11 es:

$$\begin{aligned} & P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \beta) \\ &= \left(\prod_{j=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{l=1}^{K} \theta_{j,k}^{n_{j,l,\cdot}+\alpha_k-1} \right) \left(\prod_{l=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_v-1} \right) \\ &= \left(\prod_{j=1}^{M} \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_k)^K} \prod_{l=1}^{K} \theta_{j,k}^{n_{j,l,\cdot}+\alpha_0-1} \right) \quad \left(\prod_{l=1}^{K} \frac{\Gamma(V\beta_0)}{\Gamma(\beta_0)^V} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_0-1} \right) \\ &= \left(\frac{\Gamma(K\alpha_0)^M}{\Gamma(\alpha_0)^{KM}} \prod_{j=1}^{M} \prod_{l=1}^{K} \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \quad \left(\frac{\Gamma(V\beta_0)^K}{\Gamma(\beta_0)^{VK}} \prod_{l=1}^{K} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_0-1} \right) \end{aligned} \tag{4.12}$$

Se puede marginalizar sobre las variables θ_j y φ_k para obtener $P(\mathbf{z}, \mathbf{w} | \alpha, \beta)$. Esto es posible debido a la independencia condicional de las variables que se ha explicado anteriormente. Integrando entonces sobre la ecuación 4.12 se tiene que:

$$
\begin{aligned}
&P(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta) \\
&= \int_{\theta_j} \int_{\varphi_l} \left(\frac{\Gamma\left(K\alpha_0\right)^M}{\Gamma\left(\alpha_0\right)^{KM}} \prod_{j=1}^{M} \prod_{l=1}^{K} \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \left(\frac{\Gamma\left(V\beta_0\right)^K}{\Gamma\left(\beta_0\right)^{VK}} \prod_{l=1}^{K} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_0-1} \right) \delta\theta_j \delta\varphi_l \\
&= \int_{\theta_j} \left(\frac{\Gamma\left(K\alpha_0\right)^M}{\Gamma\left(\alpha_0\right)^{KM}} \prod_{j=1}^{M} \prod_{l=1}^{K} \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \delta\theta_j \int_{\varphi_l} \left(\frac{\Gamma\left(V\beta_0\right)^K}{\Gamma\left(\beta_0\right)^{VK}} \prod_{l=1}^{K} \prod_{r=1}^{V} \varphi_{k,v}^{n_{\cdot,l,r}+\beta_0-1} \right) \delta\varphi_l
\end{aligned}
\tag{4.13}
$$

Se aprecia que estos últimos términos tienen cierta similitud con una distribución Dirichlet, motivo por el cual se puede multiplicar por 1 y obtener:

$$
\begin{aligned}
&\int_{\theta_j} \left(\frac{\Gamma\left(K\alpha_0\right)^M}{\Gamma\left(\alpha_0\right)^{KM}} \prod_{j=1}^{M} \prod_{l=1}^{K} (1)\, \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \delta\theta_j \\
&= \int_{\theta_j} \left(\frac{\Gamma\left(K\alpha_0\right)^M}{\Gamma\left(\alpha_0\right)^{KM}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)}{\Gamma\left[\sum_{l=1}^{K} \left(n_{j,l,\cdot}+\alpha_0\right)\right]} \frac{\Gamma\left[\sum_{l=1}^{K} \left(n_{j,l,\cdot}+\alpha_0\right)\right]}{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)} \prod_{l=1}^{K} \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \delta\theta_j \\
&= \frac{\Gamma\left(K\alpha_0\right)^M}{\Gamma\left(\alpha_0\right)^{KM}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)}{\Gamma\left[\sum_{l=1}^{K} \left(n_{j,l,\cdot}+\alpha_0\right)\right]} \int_{\theta_j} \left(\frac{\Gamma\left[\sum_{l=1}^{K} \left(n_{j,l,\cdot}+\alpha_0\right)\right]}{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)} \prod_{l=1}^{K} \theta_{j,l}^{n_{j,l,\cdot}+\alpha_0-1} \right) \delta\theta_j \\
&= \frac{\Gamma\left(K\alpha\right)^M}{\Gamma\left(\alpha\right)^{KM}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)}{\Gamma\left[\sum_{l=1}^{K} \left(n_{j,l,\cdot}+\alpha_0\right)\right]} (1) \\
&= \frac{\Gamma\left(K\alpha\right)^M}{\Gamma\left(\alpha\right)^{KM}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot}+\alpha_0\right)}{\Gamma\left[\left(n_{j,\cdot,\cdot}+K\alpha_0\right)\right]}
\end{aligned}
\tag{4.14}
$$

El procedimiento es análogo para la variable φ_k, por lo que al final se puede escribir la ecuación 4.13 como:

$$
\begin{aligned}
& P(\mathbf{z}, \mathbf{w}|\alpha, \beta) \\
=& \frac{\Gamma(K\alpha)^M}{\Gamma(\alpha)^{KM}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma(n_{j,l,\cdot} + \alpha_0)}{\Gamma[(n_{j,\cdot,\cdot} + K\alpha_0)]} \frac{\Gamma(V\beta_0)^K}{\Gamma(\beta_0)^{VK}} \prod_{l=1}^{K} \frac{\prod_{r=1}^{V} \Gamma(n_{\cdot,l,r} + \beta_0)}{\Gamma[(n_{\cdot,l,\cdot} + V\beta_0)]} \\
=& \frac{\Gamma(K\alpha)^M}{\Gamma(\alpha)^{KM}} \frac{\Gamma(V\beta_0)^K}{\Gamma(\beta_0)^{VK}} \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma(n_{j,l,\cdot} + \alpha_0)}{\Gamma[(n_{j,\cdot,\cdot} + K\alpha_0)]} \prod_{l=1}^{K} \frac{\prod_{r=1}^{V} \Gamma(n_{\cdot,l,r} + \beta_0)}{\Gamma[(n_{\cdot,l,\cdot} + V\beta_0)]} \\
\propto& \prod_{j=1}^{M} \frac{\prod_{l=1}^{K} \Gamma(n_{j,l,\cdot} + \alpha_0)}{\Gamma[(n_{j,\cdot,\cdot} + K\alpha_0)]} \prod_{l=1}^{K} \frac{\prod_{r=1}^{V} \Gamma(n_{\cdot,l,r} + \beta_0)}{\Gamma[(n_{\cdot,l,\cdot} + V\beta_0)]}
\end{aligned} \tag{4.15}
$$

4.3. Muestreo de Gibbs colapsado para LDA suavizado

La ecuación 4.15 da la posibilidad de evaluar la verosimilitud del modelo únicamente usando las variables $\mathbf{z}$ y $\mathbf{w}$ mediante el método del muestreo de Gibbs para el LDA. Esta implementación ha sido discutida en diversos documentos como Heinrich (2008) y Griffiths y Steyvers (2004) entre otros. El término colapsado proviene del hecho de integrar sobre los parámetros θ y φ para eliminar dichas variables y de esta manera poder tener una expresión de una probabilidad condicional de una variable aleatoria dado el resto de ellas. Es decir, se desea obtener una expresión del estilo de la ecuación 4.16.

$$
P(\boldsymbol{z_{(m,n)}}|w, \boldsymbol{z_{-(m,n)}}, \alpha, \beta) = \frac{P(\boldsymbol{z_{(m,n)}}, \boldsymbol{w}, \boldsymbol{z_{-(m,n)}}, \alpha, \beta)}{P(\boldsymbol{w}, \boldsymbol{z_{-(m,n)}}, \alpha, \beta)} \tag{4.16}
$$

Donde $z_{(m,n)}$ significa que se observa la variable z en el m-ésimo documento y la n-ésima palabra. Esto significa, por la definición de la ecuación 4.2, que en el tópico $z_{(m,n)}$ se ha seleccionado el k-ésimo componente, y por consecuencia $z_{(m,n,k)} = 1$. De la misma manera, debido a la definición de la ecuación 4.2 se ha usado el v-ésimo componente por lo que $w_{(m,n,v)} = 1$. Bajo estos supuestos, la ecuación 4.16 puede ser escrita como:

$$
\begin{aligned}
&P(z_{(m,n,k)} = 1 | w_{(m,n,v)} = 1, z_{-(m,n)}, w_{-(m,n)}, \alpha, \beta) \\
=&\frac{P(z_{(m,n,k)} = 1, w_{(m,n,v)} = 1, z_{-(m,n)}, w_{-(m,n)}, \alpha, \beta)}{P(w_{(m,n,v)} = 1, z_{-(m,n)}, \alpha, \beta)}
\end{aligned} \tag{4.17}
$$

Es notorio que en la ecuación 4.17 el denominador no depende de la variable observada $z_{(m,n,k)} = 1$, por lo que esta parte es una constante y en general se puede afirmar que:

$$
\begin{aligned}
&P(z_{(m,n,k)} = 1 | w_{(m,n,v)} = 1, z_{-(m,n)}, w_{-(m,n)}, \alpha, \beta) \\
\propto&P(z_{(m,n,k)} = 1, w_{(m,n,v)} = 1, z_{-(m,n)}, w_{-(m,n)}, \alpha, \beta) \\
\propto&\frac{\prod_{l=1}^{K} \Gamma\left(n_{j,l,\cdot} + \alpha_0\right)}{\Gamma\left[\left(n_{j,\cdot,\cdot} + K\alpha_0\right)\right]} \prod_{l=1}^{K} \frac{\prod_{r=1}^{V} \Gamma\left(n_{\cdot,l,r} + \beta_0\right)}{\Gamma\left[\left(n_{\cdot,l,\cdot} + V\beta_0\right)\right]}
\end{aligned} \tag{4.18}
$$

Considerando que $n_{j,l,r} = \sum_{i=1}^{N_j} z_{j,i,l} w_{j,i,r}$. Si se ha seleccionado el tópico k y la palabra v entonces el producto $z_{j,i,k} w_{j,i,v} = 1$. Ahora defina a $n_{j,l,r}^{-(m,n)}$ de manera similar como $n_{j,l,r}^{-(m,n)} = \sum_{i=1, i \neq n}^{N_j} z_{j,i,l} w_{j,i,r}$ y $j \neq m$ entonces $n_{j,l,r} = n_{j,l,r}^{-(m,n)} + z_{j,i,k} w_{j,i,v} = n_{j,l,r}^{-(m,n)} + 1$. Esto mismo ocurre con los demás índices y sumatorias, por lo que al observar en la variable $z_{(m,n,k)}$ y reescribiendo la ecuación 4.18 se tiene que:

$$
\begin{aligned}
&P(z_{(m,n,k)} = 1 | w_{(m,n,v)} = 1, z_{-(m,n)}, w_{-(m,n)}, \alpha, \beta) \\
&\propto \prod_{l=1, l\neq k}^{K} \Gamma\left(n_{j,l,\cdot}^{z_{-(m,n)}} + \alpha_0\right) \prod_{l=1, l\neq k}^{K} \frac{\prod_{r=1}^{V} \Gamma\left(n_{\cdot,l,r}^{z_{-(m,n)}} + \beta_0\right)}{\Gamma\left[\left(n_{\cdot,l,\cdot}^{z_{-(m,n)}} + V\beta_0\right)\right]} \\
&\times \frac{\Gamma\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0 + 1\right)}{\Gamma\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0 + 1\right)\right]} \frac{\Gamma\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0 + 1\right)}{\Gamma\left[\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0 + 1\right)\right]} \\
&\propto \frac{\Gamma\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0 + 1\right)}{\Gamma\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0 + 1\right)\right]} \frac{\Gamma\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0 + 1\right)}{\Gamma\left[\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0 + 1\right)\right]} \\
&= \frac{\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0\right)}{\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0\right)\right]} \frac{\Gamma\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0\right)}{\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0\right)\right]} \\
&\times \frac{\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0\right)}{\left[\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0\right)\right]} \frac{\Gamma\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0\right)}{\Gamma\left[\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0\right)\right]} \\
&\propto \frac{\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0\right)}{\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0\right)\right]} \frac{\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0\right)}{\left[\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0\right)\right]}
\end{aligned}
\tag{4.19}
$$

Para finalizar, note que en la ecuación 4.2 y 4.15 se observa que los parámetros θ_j y φ_k solo dependen de los hiperparámetros α y β respectivamente. Esto significa que en la ecuación 4.19, las variable θ_j y φ_k comprenden los factores que dependen de α_0 y por la parte β_0 respectivamente:

$$
\theta_j = \frac{\left(n_{m,k,\cdot}^{z_{-(m,n)}} + \alpha_0\right)}{\left[\left(n_{m,\cdot,\cdot}^{z_{-(m,n)}} + K\alpha_0\right)\right]} \tag{4.20}
$$

$$
\varphi_k = \frac{\left(n_{\cdot,k,v}^{z_{-(m,n)}} + \beta_0\right)}{\left(n_{\cdot,k,\cdot}^{z_{-(m,n)}} + V\beta_0\right)} \tag{4.21}
$$

Las ecuaciones 4.20 y 4.21 pueden ser verificadas encontrando los valores esperados de dichas variables dados los parámetros. Una deducción de estos valores esperados de forma general se encuentra en Heinrich (2008), mientras que en el capítulo 4.3.1 se presenta una forma alternativa de hacerlo utilizando la notación 1 de k.

De acuerdo a las interpretaciones de Griffiths y Steyvers (2004) θ_j representa la probabilidad del tópico k dentro del documento j y φ_k la distribución de probabilidad de las palabras en el tópico k.

4.3.1. Valores esperados de los parámetros

En esta sección se encontrará una forma de aproximar los valores de los parámetros φ_k y θ_j y se verificará que las expresiones que se obtienen, corresponden a las ecuaciones 4.21 y 4.20 respectivamente.

Es sabido que estas expresiones para el caso del muestreo de Gibbs, deben ser de tal manera que no contengan a la variable de la cual se muestrea. Estas expresiones deben quedar en términos de las variables $z_{-(m,n)}$ y $w_{-(m,n)}$. Debido a esto se trabaja con la probabilidad $p(\varphi_k|z_{-(m,n)}, w_{-(m,n)})$ tratando de simplificarla usando expresiones conocidas como en la ecuación 4.22.

$$p(\varphi_k|z_{-(m,n)}, w_{-(m,n)}, \beta) \propto p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(\varphi_k|z_{-(m,n)}, \beta) \tag{4.22}$$

Del modelo de la figura 4.2 se puede apreciar que $\varphi_k \perp\!\!\!\perp$ de $z_{-(m,n)}$ dado $\emptyset$ por lo que la ecuación 4.22 se simplifica quedando como la ecuación 4.23.

$$p(\varphi_k|z_{-(m,n)}, w_{-(m,n)}, \beta) \propto p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(\varphi_k|\beta) \tag{4.23}$$

Gracias a la definición de la ecuación 4.7, a que $p(\varphi_k|\beta) \sim Dir(\beta)$ y usando la misma notación de $n_{\cdot,k,v}$, se reescribe la ecuación 4.23 quedando como en la ecuación 4.24.

$$\begin{aligned} p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(\varphi_k|\beta) &= \prod_{v=1}^{W} \varphi_{k,v}^{n_{\cdot,k,v}^{z_{-(m,n)}}} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{v=1}^{W} \varphi_{k,v}^{\alpha-1} \\ &\propto \prod_{v=1}^{W} \varphi_{k,v}^{n_{\cdot,k,v}^{z_{-(m,n)}}+\alpha-1} \end{aligned} \tag{4.24}$$

Es decir, $p(\varphi_k|z_{-(m,n)}, w_{-(m,n)}) \sim Dir(\beta')$ donde $\beta' = (n_{\cdot,k,1}^{z_{-(m,n)}} + \beta, n_{\cdot,k,2}^{z_{-(m,n)}} + \beta, \cdots, n_{\cdot,k,W}^{z_{-(m,n)}} + \beta)$.

Empleando la definición de valor esperado se tiene que:

$$\begin{aligned} E(\varphi_{k,r}|z_{-(m,n)}, w_{-(m,n)}) &= \int \varphi_k p(\varphi_k|z_{-(m,n)}, w_{-(m,n)}) d\varphi_k \\ &= \frac{\beta + n_{\cdot,k,v}^{z_{-(m,n)}}}{W\beta + n_{\cdot,k,\cdot}^{z_{-(m,n)}}} \end{aligned} \tag{4.25}$$

De forma similar que con la ecuación 4.23, se puede usar la expresión $p(\theta_j|z_{-(m,n)}, w_{-(m,n)})$ para poder encontrar el valor esperado de $\theta_j|z_{-(m,n)}, w_{-(m,n)}$ obteniendo la ecuación 4.20 antes mencionada.

4.4. Semántica y LDA

En general, el término semántica tiene que ver con el significado que cierto símbolo tiene en un contexto. En el caso de textos, la semántica es el significado que una palabra tiene dentro del contexto de la oración o del documento.

De acuerdo con Griffiths *et al.* (2007), cuando se realiza la lectura de algún documento, es necesario que en nuestra mente se almacene una variedad de conceptos. Estos, deben estar relacionados con la información entrante que se obtiene al ir leyendo y entendiendo el texto. Mediante este proceso se puede decir que se extrae la esencia de la información.

Existe una relación entre el contexto y la aparición de las palabras que pueden ser modeladas de forma adecuada mediante el uso de distribuciones de probabilidad. Así, se reduce la tarea de extracción del contexto, desde el punto de vista computacional, a tan solo estimar la distribución de probabilidad de las palabras con base a los tópicos.

Un ejemplo basado en la semántica del contexto se visualiza en la sección 4.6.1. En la figura 4.3 se aprecia el comportamiento que indica que si la palabra "banco" va acompañada de otras como "dinero" y "préstamo", hacen referencia a un contexto relacionado con instituciones bancarias. Por otro lado, cuando "banco" va acompañada de las palabras "corriente" y "río", hacen referencia a contextos de acumulaciones de agua.

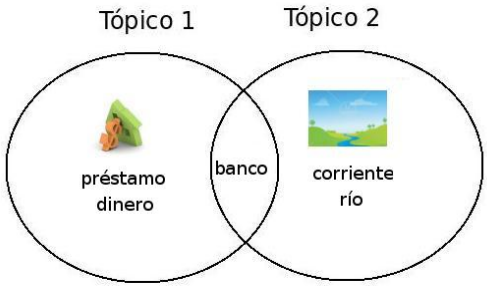


Figura 4.3: Ejemplo de la palabra "banco" manejada en diferentes contextos.

De esta forma el modelo LDA no solo agrupa palabras en tópicos, sino que también descubre la semántica de las palabras, al encontrar la mezcla adecuada con la que los documentos han sido generados. Además, el poder trabajar esta información como distribuciones de probabilidad, permite aplicar a los resultados obtenidos mediante el LDA otros procedimientos estadísticos. Esta idea favorece el estudio más a fondo del contenido y la estructura de la información.

Subsiguientemente se ejemplificará la practicidad del uso del modelo generativo basado en tópicos, para realizar aplicaciones relacionadas con la predicción de información y la recuperación automática en bases de datos de gran tamaño.

4.5. Aplicaciones del LDA

4.5.1. Similitud entre documentos y entre palabras

Como se mencionará más adelante en la sección 4.6.3.1, en ocasiones es mejor utilizar el análisis estadístico para examinar colecciones de datos de gran tamaño. Utilizando métricas de distancia entre las estimaciones de las distribuciones de probabilidad obtenidas con el modelo LDA, es posible responder algunas preguntas interesantes acerca de la estructura de los datos.

En el contexto semántico, es de suma utilidad realizar comparaciones de similitud entre entidades como lo son: las palabras y los documentos. La similitud entre palabras se da cuando estas aparecen prácticamente en los mismos tópicos. Para el caso de los documentos, la definición de similitud está basada en la obtención de distribuciones de aparición de los tópicos similares en los documentos candidatos a similitud.

Debido a que ambas definiciones trabajan con base en las probabilidades de aparición, el modelo generativo LDA proporciona un marco de trabajo intuitivo y práctico para el manejo de este tipo de problemas computacionales.

4.5.1.1. Análisis de similitud entre documentos

Como se ha explicado anteriormente, el resultado de resolver el modelo LDA en una colección de documentos consiste en obtener la distribución de probabilidad asociada con la mezcla de tópicos que genera la muestra.

Resulta natural pensar que dos documentos son similares, si de alguna manera están formados por la misma mezcla de tópicos. Luego, la forma de medir la similitud entre documentos, es cuantificando el parecido entre las distribuciones de probabilidad que representan las mezclas de tópicos para cada documento. Por lo tanto, una forma de realizar el análisis de similitud entre dos documentos d_1 y d_2, es utilizar una medida de similitud, por ejemplo, la distancia Kullback-Leibler. Se trata de medir el grado de similitud entre las distribuciones de probabilidad θ^{d_1} y θ^{d_2}, que representan las mezclas para los documentos d_1 y d_2 respectivamente.

La distancia Kullback-Leibler se define como:

$$D(p,q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j} \tag{4.26}$$

Por desgracia la distancia Kullback-Leibler no cumple con la propiedad de simetría de las métricas de distancia. Es decir, $D(p,q) \neq D(q,p)$. Por esta razón, en muchas ocasiones se utilizan otras métricas derivadas de la distancia Kullback-Leibler, como la distancia promedio de las distribuciones simétricas expresadas

en la ecuación 4.27 y la distancia de Jensen-Shannon mostrada en la expresión 4.28.

$$KL(p,q) = \frac{1}{2}\left[D(p,q) + D(q,p)\right] \tag{4.27}$$

$$JS(p,q) = \frac{1}{2}\left[D\left(p, \frac{(p+q)}{2}\right) + D\left(q, \frac{(p+q)}{2}\right)\right] \tag{4.28}$$

También se pueden interpretar estas distribuciones como simples vectores de valores reales y aplicar alguna medida de similitud geométrica como la distancia Euclidiana entre otras.

4.5.1.2. Similitud entre palabras

Generalmente la similitud entre palabras se puede medir usando las distancias KL o Js entre las distribuciones de probabilidad que se muestra en las ecuaciones 4.29.

$$\begin{aligned} \theta^{(1)} &= p(z|w_i = w_1)\ i = 1, 2, ..., V \\ \theta^{(2)} &= p(z|w_i = w_2)\ i = 1, 2, ..., V \end{aligned} \tag{4.29}$$

Otro enfoque para medir la similitud entre palabras, se basa en la relación de dependencia de la aparición de una palabra dado que en el texto ha ocurrido otra previamente, es decir, se trata de maximizar $p(w_2|w_1)$. El argumento en el que se basa esta forma de calcular similitudes, considera que para que dos palabras sean similares, deben provenir básicamente de los mismos tópicos. Debido a que los tópicos son las variables latentes (y no son observadas), la distribución de probabilidad $p(w_2|w_1)$ debe provenir de la aplicación de la regla del producto y la independencia condicional.

Esto se expresa matemáticamente como se muestra en la ecuación 4.30.

$$p(w_2|w1) = \sum_{j=1}^{K} p(w_2|z_{i,j} = 1)p(z_{i,j} = 1|w_1) \tag{4.30}$$

Si la palabra w_1 aparece, es muy probable esté relacionada con el j-ésimo tópico, esta parte es modelada por $p(z_{i,j} = 1|w_1)$. Por otro lado, ya que se seleccionó el tópico j, la probabilidad de que se elija la palabra w_2 está dada por $p(w_2|z_{i,j})$. Realizando este proceso para todos los tópicos en la colección, es posible realizar un análisis de la similitud, ya que palabras similares obtendrán valores altos de la ecuación 4.30 mientras que palabras disimiles generan valores bajos.

4.5.2. Aplicaciones en recuperación de información

La recuperación de información puede ser realizada de dos diferentes maneras.

Una forma es calcular una distribución de probabilidad de los tópicos del documento consulta q llamado θ^q. Posteriormente se calcula alguna medida de distancia entre la consulta y todos los documentos incluidos en la colección. El problema principal con esta propuesta, se encuentra en la dificultad de calcular de forma precisa dicha distribución cuando apenas se cuenta con algunas pocas palabras.

Otro enfoque consiste en encontrar el documento que maximiza la probabilidad condicional del conjunto de palabras de consulta dado cada uno de los documentos candidatos.

Según Erricson y Kintsch (1995),Kintsch (1988) y Potter (1993) este enfoque se basa en la premisa de que después de la aparición de una determinada palabra, se genera una mayor expectativa acerca de la aparición de otra, que desambigüe el contexto de la búsqueda. Si se calculara la probabilidad de que aparezcan un conjunto de palabras de consulta dentro de un documento en la colección, el efecto de desambiguación ocasionará probabilidades altas para los documentos relacionados con ellas y probabilidades bajas para los no relacionados.

Dicha probabilidad se expresa matemáticamente como en la ecuación 4.31.

$$\begin{aligned} p(q|d_i) &= \prod_{w_k \epsilon q} p(w_k|d_i) \\ &= \prod_{w_k \epsilon q} \sum_{j=1}^{K} p(w_k|z_{i,j}=1) p(z_{i,j}=1|d_i) \end{aligned} \tag{4.31}$$

En otras palabras, la meta es encontrar los tópicos que tienen una mayor probabilidad de estar relacionados con la distribución de probabilidad que definidos por las palabras de la consulta. Para esto, se hace un recorrido por cada uno de los tópicos y después se encuentra la relación de las palabras de consulta con los tópicos.

En la práctica esta técnica es la que produce mejores resultados, a diferencia de la estimación de la distribución de probabilidad de los tópicos para el documento consulta.

4.5.3. Agrupamiento de Documentos

Otra aplicación interesante que permite realizar el modelo generativo LDA es el agrupamiento de documentos.

Esta metodología en general es útil cuando se quiere encontrar alguna forma de agrupar los documentos de acuerdo a su contenido. En la sección 4.5.1.1 se

ha discutido una forma de realizar comparaciones entre documentos usando las distribuciones de probabilidad de los tópicos. El siguiente paso en el análisis es el encontrar grupos de documentos.

Los grupos de documentos pueden ser creados usando cualquier algoritmo de aglomeración, empleando como medida de distancia alguna de las expuestas en la sección 4.5.1.1. Estos grupos de documentos proporcionan información acerca de áreas que conforman la colección y sirven para diseñar sistemas de clasificación.

4.5.4. Análisis de la tendencia entre los tópicos

4.5.4.1. Tópicos de moda

En algunas colecciones de datos es importante conocer cuales tópicos son considerados de "moda" y cuales no. Como se menciona en Griffiths y Steyvers (2004) es posible realizar este tipo de análisis si la base de datos cuenta con suficiente información. Por ejemplo, en bases de datos de diversas ciencias, se pueden agrupar por áreas del conocimiento y entonces estudiar las proporciones de aparición de cada uno de los tópicos dentro de las áreas. Es natural pensar que los tópicos con mayor probabilidad son los tópicos de moda acerca de los cuales se discute con mayor frecuencia.

4.5.4.2. Progresión de tópicos de interés por períodos de tiempo

Al igual que el estudio de los tópicos, también es posible realizar un análisis de progresión por años. Por medio de este análisis se estudia la evolución del grado de interés de los tópicos en un período de tiempo, e inclusive permite hacer predicciones acerca de futuros tópicos de interés.

No obstante, una desventaja de este tipo de análisis, consiste en que requiere que el número de tópicos se mantenga fijo a lo largo del período de estudio, impidiendo que aparezcan nuevos tópicos.

Otro inconveniente importante se encuentra en el análisis progresivo. Al ser progresivo se requiere que nueva información sea agregada cuando se obtenga. Sin embargo, el método del muestreo de Gibbs al ser un proceso estocástico no proporciona un orden específico para la aparición de los tópicos. Por lo tanto, encontrar la correspondencia adecuada entre los tópicos para cada corrida en un período de tiempo, puede resultar en otra tarea computacionalmente costosa.

4.6. Experimentos

En esta sección, se muestra un conjunto de experimentos que intentan explicar y demostrar el funcionamiento del modelo de gráficas LDA. La mayoría de los

experimentos que se presentan a continuación están basados en el análisis de texto, esto es debido a que este modelo surgió inicialmente como una propuesta para el estudio de grandes corpus de textos. También se presenta un ejemplo sintético y didáctico basado en imágenes y patrones de franjas.

4.6.1. Experimento sintético 1

Consiste en replicar un experimento didáctico sugerido en Steyvers y Griffiths (2007) y cuyos datos de prueba fueron tomados de `http://alias-i.com/lingpipe/demos/tutorial/cluster/read-me.html`. Se generan de manera sintética un conjunto de 16 documentos cada uno con sus respectivas 16 palabras, basado en un diccionario de 5 elementos cuya distribución se muestra en la tabla 4.1.

	río	corriente	banco	dinero	préstamo
Tópico 1	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
Tópico 2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0

Cuadro 4.1: Distribución de probabilidad de las palabras sobre los 2 tópicos.

Como resultado del muestreo se obtiene la colección de datos que se observa en la tabla 4.2. En esta, cada fila representa la distribución de las palabras dentro de cada documento. Cada columna indica la frecuencia de aparición de las palabras agrupadas por tópico. Igualmente se presentan los totales de aparición de cada palabra separado por tópicos en el corpus, así como, las proporciones de aparición de un tópico para cada documento.

	Tópico 1				Tópico 2			
No Doc	banco	dinero	préstamo	Prop	río	corriente	banco	Prop
1	4	6	6	1	0	0	0	0
2	5	7	4	1	0	0	0	0
3	7	5	4	1	0	0	0	0
4	7	6	3	1	0	0	0	0
5	7	2	7	1	0	0	0	0
6	9	3	4	1	0	0	0	0
7	4	6	5	0.937	1	0	0	0.062
8	6	4	3	0.812	1	2	0	0.187
9	3	4	2	0.562	1	3	3	0.437
10	5	1	4	0.625	2	3	1	0.375
11	6	3	1	0.625	2	3	1	0.375
12	0	1	0	0.0625	3	6	6	0.937
13	0	0	1	0.0625	6	3	6	0.9375
14	0	0	0	0	2	8	6	1
15	0	0	0	0	4	7	5	1
16	0	0	0	0	5	7	4	1
Total	63	48	44		27	42	32	

Cuadro 4.2: Distribución de probabilidad de las palabras sobre los 2 tópicos.

Los resultados esperados deben ser parecidos a los presentados en la tabla 4.2. Al correr el algoritmo del muestreo Gibbs colapsado para el LDA usando como parámetro $\alpha = 1$,$\beta = 0.01$ y $k = 2$, con 100000 iteraciones como criterio de convergencia de la cadena, se obtienen los resultados que se aprecian en las tablas 4.3 y 4.4.

	Tópico 1			Tópico 2		
	banco	dinero	préstamo	río	corriente	banco
Total	64	48	44	27	42	31

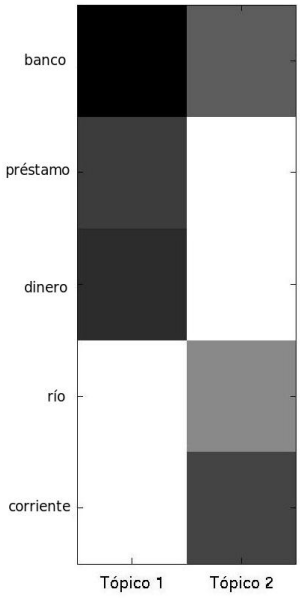
Cuadro 4.3: Frecuencia de aparición de las palabras por tópico.

	Entrada		Salida	
No Doc	Prop. T1	Prop. T2	Prop. T1	Prop. T2
1	1	0	1	0
2	1	0	1	0
3	1	0	1	0
4	1	0	1	0
5	1	0	1	0
6	1	0	0.937	0.062
7	0.937	0.062	0.75	0.25
8	0.812	0.187	0.687	0.3125
9	0.562	0.437	0.625	0.375
10	0.625	0.375	0.625	0.375
11	0.625	0.375	0.625	0.375
12	0.0625	0.937	0.312	0.687
13	0.062	0.937	0.125	0.875
14	0	1	0	1
15	0	1	0	1
16	0	1	0.0625	0.9375

Cuadro 4.4: Comparación entre las proporciones verdaderas y las calculadas.

Como se puede ver en la tabla 4.4, a pesar de contar con pocos datos (solo 16 documentos) se encuentran las proporciones (θ_j y φ_k) que aproximan a las verdaderas. A pesar de que en algunos casos el algoritmo logra descubrir de forma precisa las proporciones de los tópicos en los documentos y de las palabras en el corpus, en la mayoría de los casos existen pequeñas diferencias.

Las frecuencias de aparición de las palabras aparecen en los cuadros 4.5 y 4.6. El 4.5 muestra gráficamente la frecuencia de aparición de las palabras para cada uno de los tópicos. En el 4.6 se observan 4 documentos de la colección de 16, que demuestran que en ambos tópicos la palabra "banco" aparece en múltiples ocasiones.



Cuadro 4.5: Proporciones de aparición en la colección.

Doc1	Doc2	Doc15	Doc16
banco	préstamo	río	corriente
préstamo	dinero	corriente	río
dinero	dinero	corriente	río
préstamo	préstamo	corriente	banco
préstamo	banco	río	corriente
dinero	banco	corriente	corriente
banco	dinero	corriente	corriente
préstamo	banco	banco	corriente
banco	dinero	banco	banco
préstamo	préstamo	banco	río
préstamo	dinero	banco	río
dinero	dinero	río	corriente
banco	banco	río	banco
dinero	préstamo	corriente	río
dinero	banco	banco	corriente
dinero	dinero	corriente	banco

Cuadro 4.6: Muestra de 4 documentos.

La precisión de esta estimación podría incrementar si se aumenta el número de datos, es decir, se aumenta la cantidad de documentos en el corpus al igual que el número de palabras que cada documento contiene. De esta manera, las palabras tendrán un menor peso y los errores cometidos en consecuencia afectarán menos. Sin embargo, es importante señalar que a diferencia de muchos algoritmos, el muestreo de Gibbs no mejora los resultados si se le permite un mayor número de iteraciones después de la convergencia. Una vez que el algoritmo comienza a muestrear de la distribución estacionaria, la cadena ha convergido, y lo único que se requiere, es tomar suficientes valores como para estimar de forma adecuada los parámetros.

Cabe señalar que los parámetros tales como la dimensión del vector α y valor de los vectores α y β en este experimento fueron propuestos de manera arbitraría utilizando el conocimiento a priori que se tiene de los datos. En la práctica el uso de dicho conocimiento a priori es permitido, pero en pocas ocasiones se cuenta con dicha información.

Comúnmente el conocimiento acerca del valor adecuado para dichos parámetros no va más allá de saber cómo manejar la concentración de las palabras y de los tópicos (parámetros θ_j y φ_k), manipulando el valor de α y β. La regla en general, se deriva de las propiedades de distribución Dirichlet y permite decidir como concentrar las distribuciones de probabilidad multinomiales. Por ejemplo, cuando los valores del parámetro α cumplen que $0 < \alpha_0 << 1$ entonces las

distribuciones multinomiales concentrarán las probabilidades en uno o a lo más un conjunto reducido de tópicos. Por el contrario valores cercanos a 1 generan distribuciones multinomiales que mezclaran todos los posibles tópicos, mientras que valores mayores que 1 preferirían mezclas de ciertos conjuntos siendo más homogéneos.

Por otro lado, la estimación de la dimensión del vector α es otro problema de selección de modelo que se soluciona según lo sugerido en Griffiths y Steyvers (2004) bajo el apartado de "selección de modelo". La estrategia consiste en graficar la verosimilitud dada por $p(w|z, K)$, para un valor determinado de K e ir variando el valor de K hasta encontrar el máximo. Dicha probabilidad refleja el nivel de certidumbre con la que los datos observados w y los tópicos obtenidos z se explican mediante el parámetro K con valor k_0. Entonces, el máximo reflejará la mayor probabilidad de que los datos observados provengan de dicho parámetro K.

En algunas ocasiones por cuestiones numéricas se suele calcular en lugar del valor de $p(w|z, K)$ el valor de $\log p(w|z, K)$. Debido a que $p(w|z, K)$ contiene implícitamente el cálculo de funciones Gamma, las cuales toman valores grandes muy rápido. Un ejemplo de este procedimiento se presentará en el siguiente sección.

4.6.2. Experimento sintético 2

El objetivo de este experimento, es apreciar visualmente la forma en que el modelo LDA trabaja. Para esto se ha seguido el ejemplo de imágenes con barras que se muestra en Griffiths y Steyvers (2004). Se generaron un conjunto de 5,000 documentos de prueba, tomando muestras de los tópicos representados por imágenes de tamaño 5 x 5 píxeles y cada uno de los píxeles es una palabra del diccionario. Las imágenes que se usan como tópicos consisten en un patrón de franjas horizontales y verticales como las que se muestran en la figura 4.4.

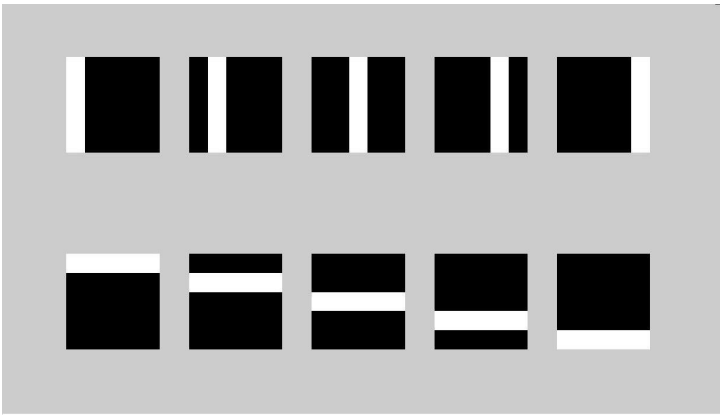


Figura 4.4: Tópicos desde los cuales fueron generadas los patrones de franjas.

Estos patrones de franjas definen las distribuciones de probabilidad de las palabras sobre los tópicos y juegan el papel de la matriz φ_k del modelo gráfico del LDA presentado en la figura 4.2. Por otro lado, los documentos se forman al muestrear los tópicos conforme a una distribución multinomial obtenida a través de la distribución Dirichlet con parámetros $\alpha_i = 1$ según el algoritmo 4.1.1. Posteriormente, se muestrea un píxel de dicho tópico uniformemente, repitiendo este procedimiento tantas veces como número de palabras (en este caso píxeles) se requieran en cada documento. Conforme a lo anterior, se observa que cada uno de los documentos tendrán un tamaño fijo de 100 elementos por lo que de acuerdo al algoritmo 4.1.1 se puede establecer $N_j = 100$ para $j = 1, 2...200$.

Es importante no confundir el tamaño del diccionario con el tamaño del documento. El diccionario está formado de 25 elementos y lo definen las dimensiones de la imagen, mientras que el tamaño de cada documento es de 100 elementos y representa el número de muestras tomadas. Así, se puede construir una matriz de ocurrencias que representaría a una imagen de 5 x 5 píxeles, donde cada píxel de la imagen es una palabra y su intensidad representará la frecuencia de ocurrencia de dicha palabra sobre un documento. Este muestreo, en representación de escala de grises se observa en la figura 4.5.

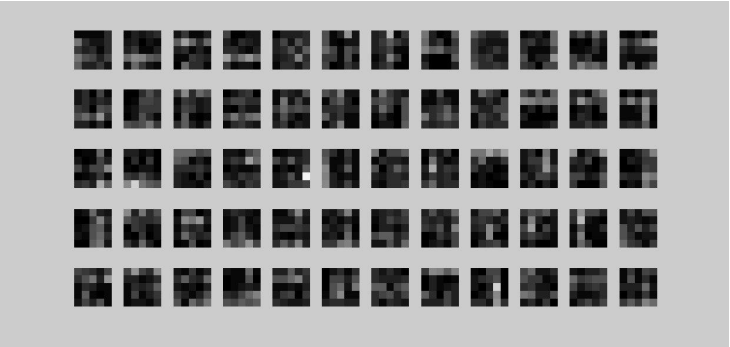


Figura 4.5: Un subconjunto de datos de entrenamiento usados para aprender los parámetros del modelo.

El objetivo principal de este experimento es encontrar de forma adecuada la distribución que los tópicos tienen sobres las palabras, es decir, como salida se debe poder visualizar algo similar a lo presentado en la figura 4.4. Los resultados obtenidos en diferentes momentos de la corrida del algoritmo se exponen en la figura 4.6.

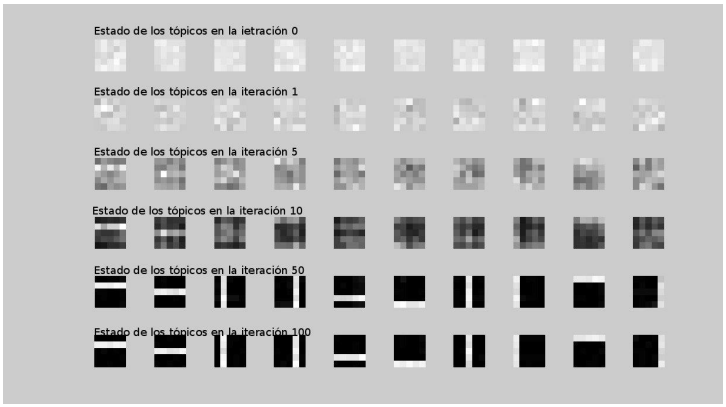


Figura 4.6: Evolución de los tópicos en diferentes iteraciones.

Se aprecia que conforme el número de iteraciones va en aumento, la calidad del resultado lo hace también; ya que en las primeras corridas los resultados no sugerían absolutamente nada mientras que a partir de la iteración 50 ya se tienen resultados aceptables. En la iteración 100 el algoritmo presenta resultados muy próximos, pudiendo encontrar el patrón de franjas que generó inicialmente los datos. Esto ocurre debido al comportamiento del muestreo de Gibbs, en el cual cuando se tienen pocas iteraciones, el muestreo aún diverge de la verdadera distribución de probabilidad. Sin embargo, conforme aumenta el número de iteraciones, el muestreo eventualmente va convergiendo a dicha distribución, por lo que la calidad de la estimación de los parámetros eventualmente mejora también. Este comportamiento puede ser verificado en la figura 4.7, donde además se muestra la comparación de la eficiencia entre el comportamiento del muestreo de Gibbs con el método de estimación "Variacional Bayesiano".

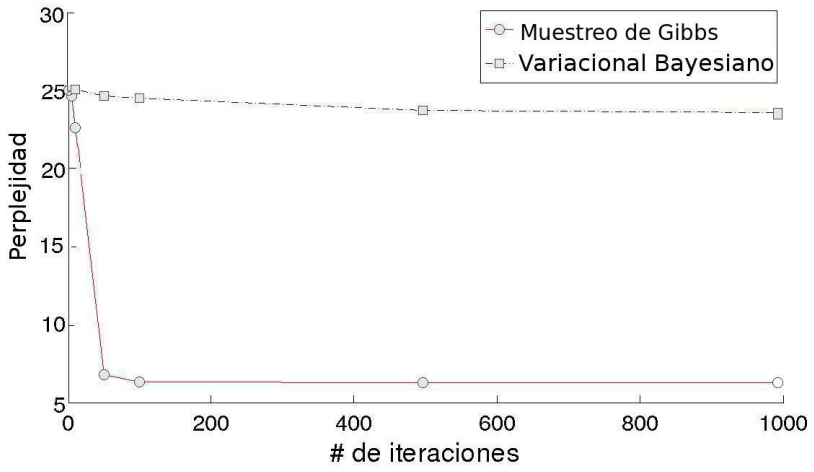


Figura 4.7: Evolución de la estimación con respecto al número de iteraciones.

En la figura 4.7 se grafica el término perplejidad utilizado para medir la calidad de los modelos estadísticos de lenguaje Natural y sugerido inicialmente en Manning y Schuetze (1999). La perplejidad se define como:

$$perp(\mathbf{w}) = exp\left(-\frac{\sum_{j=1}^{D}\log p(w_j|\varphi)}{\sum_{j=1}^{D}N_j}\right) \tag{4.32}$$

En general, se puede decir que valores de perplejidad más bajos reflejan un mejor desempeño del modelo sobre los datos. En la figura 4.7 muestra la perplejidad contra el número de iteraciones. Los datos usados son producidos al correr 10 veces el mismo algoritmo y promediar el valor obtenido. Para el caso del algoritmo del muestreo de Gibbs se utilizaron valores iniciales aleatorios para la $\boldsymbol{\varphi}$ y $\boldsymbol{\alpha}$, mientras que para el algoritmo varacional Bayesiano el valor inicial de $\boldsymbol{\varphi}$ fue obtenido muestreando de la Dirichlet con el valor del parámetro $\beta = 5$ y también $\alpha_i = 1$. Los algoritmos fueron medidos probados para 1, 5, 50, 100, 500 y 1000 iteraciones respectivamente.

Los resultados de la gráfica ejemplifican cómo el algoritmo del muestreo de Gibbs mejora su calidad de estimación cuando aumentan las iteraciones. Cabe señalar que a partir de 50 iteraciones ya se obtiene un buen valor de aproximación. El usar un número mayor de iteraciones a 100, no presenta una mejora significativa.

Para los resultados de método Variacional Bayesiano se observan dos de sus principales inconvenientes:

- La selección del punto inicial.
- Facilidad de Convergencia a mínimos locales.

Estos dos problemas están relacionados entre ellos, ya que una mala selección de puntos iniciales puede llevar a una convergencia prematura hacia un mínimo local, como se aprecia en la figura 4.7. Se aprecia también en dicha figura que la convergencia es muy lenta, ya que a diferencia del muestreo de Gibbs el algoritmo Variacional Bayesiano sigue disminuyendo entre 500 y 1000 iteraciones aunque lo hace de manera muy lenta.

4.6.2.1. Selección de modelo

Ahora bien es importante mencionar que los parámetros elegidos para la resolución de los modelos usando el algoritmo LDA con muestreo de Gibbs suavizado,

fueron escogidos mediante el proceso de selección de modelo que se describe al final del ejemplo anterior. Recordando la estrategia se trata de estimar la probabilidad $p(w|z, K)$.

Debido a que esta estimación depende de la variable aleatoria z, la cual se calcula a través de un proceso de muestreo estadístico, en el cual los valores obtenidos de z serán diferentes en cada instancia y momento del proceso. Esto genera variaciones en la medición, pero según lo expuesto en Kass y Raferti (1995), se sugiere tomar varias muestras de los cálculos de esta probabilidad, de tal forma que las mediciones pertenezcan a la misma cadena y a nuevas cadenas inicializadas con diferentes semillas. Consecutivamente se calcula el valor medio de dicha probabilidad usando la media armónica definida por:

$$\hat{p}(w|z, K) = \left[\frac{1}{M} \sum_{i=1}^{M} p(w_i|z_i, K)^{-1} \right]^{-1} \tag{4.33}$$

Mediante este procedimiento y tomando suficiente muestras se puede garantizar encontrar una buena aproximación a este valor, ya que el proceso ha sido generalizado al realizar el cambio de semilla en cada cadena. Debido que todas las cadenas deben converger a la misma distribución estacionaria, la ecuación 4.33 presenta una forma estable de aproximar el verdadero valor de $p(w|z, K)$.

La forma adecuada de poder definir el modelo a usar se basa en la elección de un valor adecuado de los parámetros. Esto se logra analizando las gráficas obtenidas al hacer variar los parámetros del modelo. Para esto se fija un valor de α y se grafican diferentes valores de β haciendo variar los valores de K. Subsiguientemente se selecciona el valor o valores del parámetro que según a juicio de los expertos en los datos, funcione de manera más adecuada y luego se hace variar el valor de α con el valor de la β seleccionada y múltiples valores de k.

La idea de realizar este procedimiento es tratar de identificar algún patrón que sugiera algún intervalo de valores de los parámetros, en los cuales se analizen de forma cualitativa la calidad de los resultados.

La figura 4.8 demuestra este procedimiento fijando los valores de $\alpha = 1$. Se observa que en las cuatro gráficas el valor máximo está alrededor de los 10 tópicos.

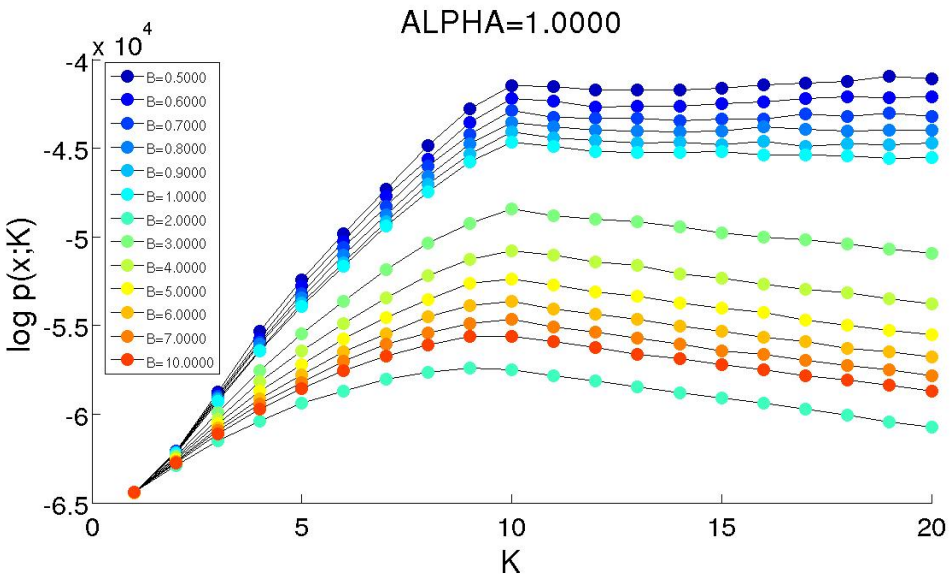


Figura 4.8: Diversos valores de β y $\alpha = 1$ para la gráfica de la selección de modelo.

En el caso específico de este ejemplo, encontrar el valor adecuado para el parámetro β es la meta. En las gráficas de la figura 4.8 se puede ver el cálculo de $\hat{p}(w|z,k)$ para diversos valores de β.

Es claro que para todas las gráficas el punto $K = 10$ es un máximo local. Esto sugiere que dicho punto es un valor interesante para ser revisado. También se nota que para valores $\beta < 1$, el comportamiento de dicha curva es bastante inestable.

Este comportamiento se debe a que en realidad la probabilidad a maximizar para esta curva debería ser aquel dado por $p(w, z, K) = p(w|z, K)p(z, K)$. Pero debido a que la distribución $p(z, K)$ es desconocida, entonces se opta por trabajar únicamente con la verosimilitud $p(w|z, K)$.

Por el contrario, cuando la $\beta \geq 1$ se aprecia un comportamiento más suave con la gráfica. En ellas, además se observa que cuando $\beta \geq 1$ ya solo existe un máximo global, el cual está dado por los valores de $K = 10$. De aquí en adelante una inspección cualitativa permitirá escoger un valor adecuado para los parámetros.

Este método puede en ocasiones llegar a ser malo, ya que algunos de los cálculos son muy complicados en términos de tiempo computacional; esto porque se requiere correr múltiples veces el muestreo de Gibbs con diferentes parámetros. El proceso de selección de modelo es como se aprecia, es un proceso bastante complejo. Una alternativa a este proceso de selección de modelo está en el uso métodos de estadística bayesiana no paramétrica, como los descritos en Muller y Quintana (2004).

La selección adecuada del método a aplicar para elegir el modelo, en muchas ocasiones dependerá únicamente del conocimiento a priori que se tenga del conjunto de datos a analizar.

4.6.3. Experimento de texto con datos reales

4.6.3.1. Base de datos de NIPS

Este ejemplo muestra el análisis de la base de datos conformada por un conjunto de artículos de la fundación llamada "Neural Information Processing System". La base de datos de este experimento consta de una colección de 1740 documentos que contienen un total de 2,301,375 palabras con un vocabulario de 13,649 términos diferentes.

Como se observa, el tamaño de esta base de datos imposibilita un análisis cualitativo por simple inspección. Con la ayuda de técnicas estadísticas y probabilísticas es posible obtener mediciones que resumen la estructura de la información, siendo estas usadas como base para este tipo de análisis.

Una muestra de esto se aprecia al observar el resultado obtenido de correr el algoritmo del LDA que se presenta en la tabla 4.7.

TOP 1	**0.03264**
units	0.08944
hidden	0.0618
unit	0.05269
layer	0.04984
network	0.0414
input	0.04005
output	0.03544
weights	0.0306
net	0.02012
propagation	0.01653
back	0.01552
training	0.01428
networks	0.01293
figure	0.0114
test	0.00708

TOP 13	**0.01622**
classification	0.07065
class	0.06749
classifier	0.0442
classes	0.03021
classifiers	0.02583
pattern	0.01749
feature	0.01416
nearest	0.01352
decision	0.01325
problem	0.0101
training	0.00999
classified	0.00986
vectors	0.00972
patterns	0.00905
neighbor	0.00863

Top 24	**0.01720**
mixture	0.04439
data	0.04404
likelihood	0.0364
em	0.03358
density	0.02338
gaussian	0.02265
parameters	0.0225
log	0.01821
experts	0.01799
maximum	0.01484
mixtures	0.01371
estimation	0.01363
set	0.01224
expert	0.01151
missing	0.01109

TOP 25	0.02579
algorithm	0.12481
algorithms	0.04129
convergence	0.02938
step	0.02219
time	0.0167
results	0.01535
update	0.01335
iteration	0.01246
iterations	0.0111
rate	0.00862
converge	0.00855
problem	0.00833
problems	0.00818
line	0.00768
fixed	0.00739

TOP 43	0.02079
matrix	0.05931
vector	0.05208
vectors	0.03814
linear	0.03414
space	0.01976
analysis	0.01857
dimensional	0.01669
component	0.01519
principal	0.01419
components	0.01308
pca	0.013
projection	0.01288
matrices	0.01259
diagonal	0.00996
transformation	0.00963

TOP 50	0.02203
bayesian	0.03015
distribution	0.02564
gaussian	0.02413
prior	0.02069
posterior	0.02047
data	0.01585
parameters	0.01184
variables	0.01176
sampling	0.01113
approximation	0.01072
inference	0.01007
belief	0.00946
carlo	0.00934
monte	0.00924
log	0.00909

Cuadro 4.7: 6 tópicos tomados de la base da datos de NIPS con la primeras 15 palabras más frecuentes.

Estos tópicos propuestos en la tabla 4.7 tienen una notoria caracterización. Por ejemplo, analizando el tópico 25 se puede ver que dichas palabras están relacionadas con temas de algoritmos numéricos computacionales, al igual que el tópico 1 que hace referencia a redes neuronales y el tópico 50 al de inferencia Bayesiana. Por supuesto que los tópicos solo tienen coherencia si es posible identificar la temática a la cual pertenecen las palabras, cosa que en muchas ocasiones solo es hecho por expertos en el las áreas de estudio.

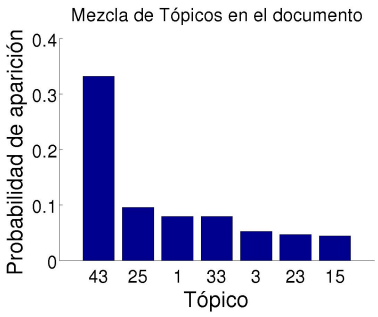
Un análisis cualitativo de los documentos también se presenta en la figura 4.9, en la cual se transcriben dos párrafos del resumen de un artículo en esta colección de datos, pero se espera que el resto del documento se comporte de manera similar. En dicha figura se observa la forma en la que la mezcla de tópicos da lugar a parte de este documento. Para este propósito se muestra cada palabra perteneciente a un tópico agrupadas por colores y también se etiqueta con el número del tópico escribiéndolo en la esquina superior derecha.

The Singular[43] Value Decomposition[43] (SVD) is an important tool for linear[43] algebra[43] and can be used to invert[43] or approximate[25] matrices[43]. Although many authors use "SVD" synonymously with "Eigen[43]- vector[43] Decomposition" or "Principal[43] Components[43] Transform", it is important to realize that these other methods[25] apply only to symmetric matrices [43], while the SVD can be applied to arbitrary nonsquare matrices[43]. This property is important for applications to signal transmission and control.
I propose two new algorithms[25] for iterative[25] computation[25] of the SVD given only sample inputs and outputs from a matrix[43]. Although there currently exist many algorithms[25] for Eigenvector[43] Decomposition[43] (Sanger 1989, for example), these are the first true sample based SVD algorithms[25].

Figura 4.9: Abstract del artículo titulado "Two Iterative Algorithms for Computing the Singular Value Decomposition from Input/Output Samples" escrito por Terence D. Sanger y obtenido de la base de datos de NIPS.

Los resultados del análisis cualitativo propuesto en la figura 4.9, se puede ver reforzado mediante el uso de un análisis cuantitativo basado en las muestras de la variable latente z arrojadas por el método del muestreo de Gibbs. Este análisis numérico en muchas ocasiones aumenta el conocimiento que se tiene acerca de dicho documento, ya que presenta un resumen sobre los tópicos con los que se genera el documento. En la tabla 4.8 se realiza dicho análisis cuantitativo y los primeros 2 tópicos con mayor probabilidad de aparición, son precisamente los mismos que se han resaltado en el análisis cualitativo. En muchas ocasiones, estos análisis cuantitativos son más eficientes al ser más prácticos para ser trabajados usando herramientas computacionales.

Tópico	Proporción
43	**0.332164**
25	**0.095906**
1	0.079532
33	0.079532
3	0.052632
23	0.046784
15	0.044444



Cuadro 4.8: Valores de las proporciones de la mezcla de tópicos y su gráfica de barras.

El cuadro 4.9 muestra el análisis de similitud para los tópicos 41 y 50 dado que en la consulta se ha introducido la palabra "distribución". El proceso pa-

ra la obtención de la probabilidad de la fórmula 4.31 implica hacer los cálculos como los del cuadro 4.9 para cada uno de los tópicos y para todos los documentos. Consecutivamente se retornan los documentos que hayan obtenido mayores probabilidades.

TOP 41	
probability	0.0028401012
information	0.0014219234
information	0.0007135368
sample	0.0006915314
random	0.000669526
density	0.000636752
distributions	0.0005777588
log	0.0005749496
theory	0.0005599672

TOP 50	
bayesian	0.000773046
gaussian	0.0006186932
prior	0.0005304916
posterior	0.0005248508
data	0.000406394
parameters	0.0003035776
variables	0.0003015264
sampling	0.0002853732
approximation	0.0002748608

Cuadro 4.9: Probabilidad de las palabras pertenecientes a los tópicos 43 y 50 dado que ocurre la palabra **Distribution**.

4.6.4. Base de datos de WormsBase

Esta base de datos fue creada con la información contenida en los resúmenes de diversos artículos, cuyo tema principal está basado en las características y condiciones de vida de los nematodos (organismos de vida similares a los gusanos). La base de datos a procesar fue obtenida de `ftp://ftp.wormbase.org/pub/wormbase/misc/literature/2007-12-01-wormbase-literature.endnote.gz` y procesada para tener una colección de $24,484$ documentos, que contienen un total de $4,060,908$ palabras y definen un diccionario de $74,538$ diferentes términos libres de *Stop Words*.

Algunos de los tópicos obtenidos como resultado de correr el algoritmo del muestreo de Gibbs usando un parámetro de concentración de las palabras $\beta = 0.01$ y para el parámetro de concentración de los tópicos de $\alpha = 0.1$.

TOP 4	0.01839
membrana	0.0466
proteins	0.03567
protein	0.02045
cuticle	0.01611
transport	0.01591
collagen	0.0131
surface	0.01204
extracellular	0.01064
plasma	0.0106
pat	0.00966

TOP 9	0.0193
muscle	0.08457
gfp	0.07715
body	0.05146
pharyngeal	0.03162
expression	0.03076
muscles	0.02886
expressed	0.02781
wall	0.02604
cells	0.01757
pharynx	0.01704

TOP 22	0.02885
gene	0.04531
sequence	0.03636
region	0.03479
cdna	0.02646
kb	0.02143
genomic	0.01961
sequences	0.01672
clones	0.01629
genes	0.01527
pcr	0.01469

TOP 33	0.0210
family	0.06132
gene	0.042
genes	0.03563
conserved	0.03371
elegans	0.0309
drosophila	0.03047
proteins	0.02894
human	0.0267
members	0.02614
related	0.02236

TOP 44	0.0245
mutations	0.10588
alleles	0.04357
mutation	0.03904
phenotype	0.03426
allele	0.02447
gene	0.02324
suppressors	0.01927
genes	0.01838
dominant	0.01812
function	0.0159

TOP 49	0.02049
amino	0.04547
acid	0.03826
n	0.02226
sequence	0.02029
protein	0.02003
acids	0.0191
residues	0.01596
activity	0.01586
alpha	0.0151
terminal	0.0136

Cuadro 4.10: 6 tópicos tomados de la base da datos de WormBase con la primeras 15 palabras más frecuentes.

Se aprecia que las palabras agrupadas en los tópicos tienen cierta estructura, que permite entender la temática de los tópicos aún sin tener conocimiento previo de la información que contienen. El algoritmo del muestreo de Gibbs mediante logra encontrar la forma en la que se distribuyen las palabras para agruparlas en tópicos y también logra mostrar la forma en la que los documentos han sido generados.

La figura 4.10 resume la información acerca de la mezcla de tópicos para el caso de 100 documentos escogidos mediante un muestreo aleatorio distribuido uniformemente en el corpus. Para analizar la mezcla de documentos en la colección basta con fijarse en la columna correspondiente al documento y ver cuales son los tópicos que tienen colores más obscuros. Por ejemplo, la figura 4.11 muestra gráficamente la forma en la que está mezclado el documento titulado "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin", el cual ocupa la columna 19 de la figura 4.10. Esta figura enseña que este documento

está conformado de los tópicos 19 y 39 que representan los picos más grandes en las gráficas.

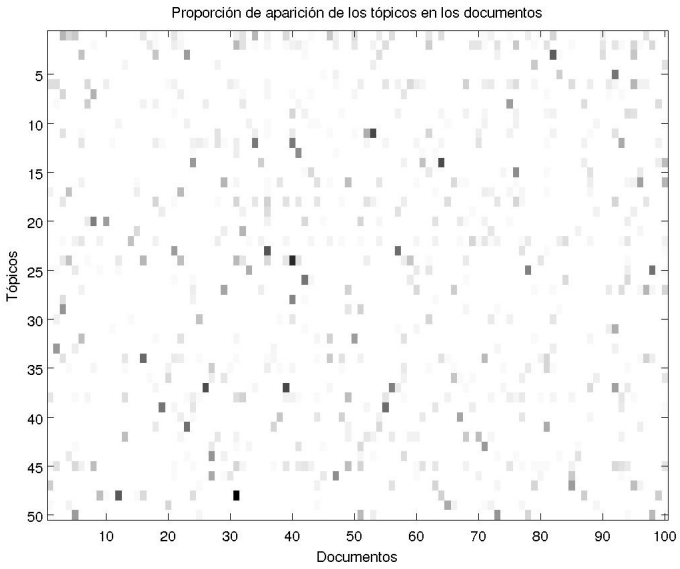


Figura 4.10: Proporción de aparición de los tópicos en una muestra aleatoria de 100 documentos de la colección.

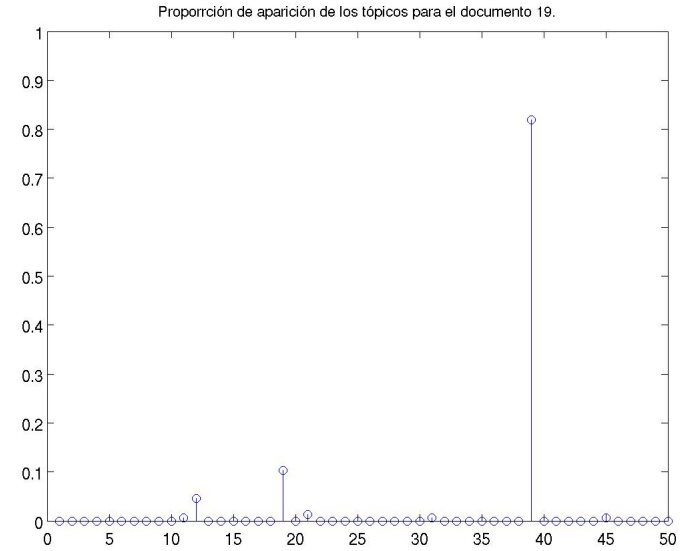


Figura 4.11: Mezcla de tópicos para el documento titulado "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin".

La tabla 4.11 presenta las primeras 10 palabras de los tópicos 19 y 39 que para el caso específico del documento "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin" tienen proporciones de aparición de 0.1032 y 0.8194 respectivamente.

TOPIC 19	0.1032
1	0.04597
pathway	0.043
let	0.03688
signaling	0.03148
ras	0.02926
function	0.02864
60	0.01811
2	0.01397
phenotype	0.01391
downstream	0.01366

TOPIC 39	0.8194
neurons	0.0616
sensory	0.03872
neuron	0.01656
osm	0.01322
chemotaxis	0.01297
response	0.01222
mutants	0.01198
cilia	0.01061
nhr	0.0106
awc	0.01058

Cuadro 4.11: Primeras 5 palabras de cada tópico que forma al documento 19 de la figura 4.11.

La figura 4.12 expone el análisis del resumen etiquetado automáticamente, tomando la salida de la variable aleatoria z que proporciona el algoritmo del muestreo de Gibbs. Es claro que en su mayoría las palabras son etiquetadas bajo el tópico 39, en algunas ocasiones también aparecen el tópico 19, mientras que ocasionalmente aparecen también los tópicos 12 y 45.

Asymmetric[19] *cell*[19] *division*[19] *is a fundamental*[19] *process*[19] *that produces*[19] *celler*[19] *diversity*[19] *during development*[19]. *In C*[19]. *elegans*[12], *asymmetric*[19] *divisions*[19] *of certain*[19] *blast*[19] *cells*[19] *including*[19] *T*[19] *blast*[19] *cell*[39] *are regurated*[19] *by lin*[39]*-17*[19]*/frizzled*[19] *and lin*[39]*-44*[19]*/wnt*[19]. *It has been proposed*[39] *that LIN*[19]*-44*[19] *signal*[19] *which acts*[19] *through LIN*[39]*-17*[19] *receptor*[19],*provides*[45] *polarity*[19] *to cells*[19] *that undergo*[19] *asymmetric*[19] *division*[19]. *To make*[19] *clear*[12] *this model*[19], *we expressed*[19] *lin*[39]*-44*[19] *ectopically*[19], *examined effects asymmetric*[19] *cell*[19] *division*[19]. *In normal*[19] *development*[19], *the anterior*[19] *daughter*[19] *of T*[19] *cell*[19] *produces*[19] *hypodermal*[19] *cells*[19], *and the posterior*[19] *daughter*[19] *produces*[19] *neural*[19] *cells*[19]. *In lin*[19]*-44*[19] *mutants*[19], *howerver, the anterior*[19] *daughter*[19] *produces*[19] *neural*[19] *cells*[19], *and the posterior*[19] *daughter*[19] *produces*[19] *hypodermal*[39] *cells*[19].

Figura 4.12: Extracto del resumen del artículo titulado "Regulation of cell polarity and asymmetric cell division by lin-44wnt and wrm-1-catenin" etiquetados de forma automática.

En la figura 4.13 se resume el análisis de frecuencia de aparición de las palabras en los tópicos únicamente para las 2 palabras más recurrentes, ya que como se ha mencionado anteriormente, el diccionario de palabras para este ejemplo consta de $74,538$ valores que difícilmente se podrían presentar todas juntas.

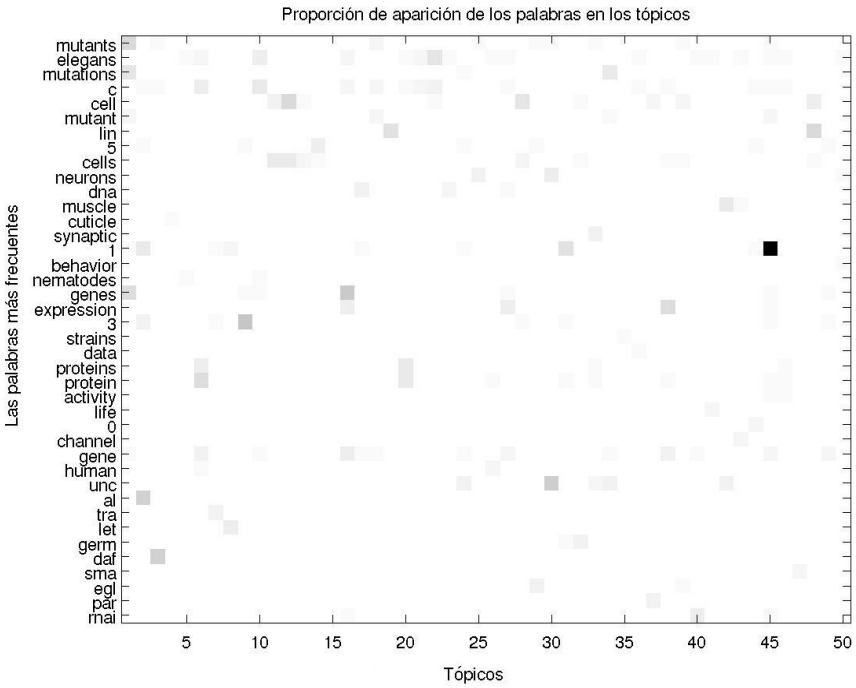


Figura 4.13: Proporción de aparición de la primera palabra de cada tópico.

Otra aplicación del modelo LDA, como se ha mencionado anteriormente, es la recuperación automatizada de documentos. Por ejemplo, dado un determinado documento consulta, se intenta recuperar otros documentos que tengan distribuciones de probabilidad parecidas tal y como se menciona en la sección 4.5.1.1. El resultado de aplicar esta metodología, usando como medida de similitud la expresión 4.27 y como documento consulta el 19, se presenta en la tabla 4.12.

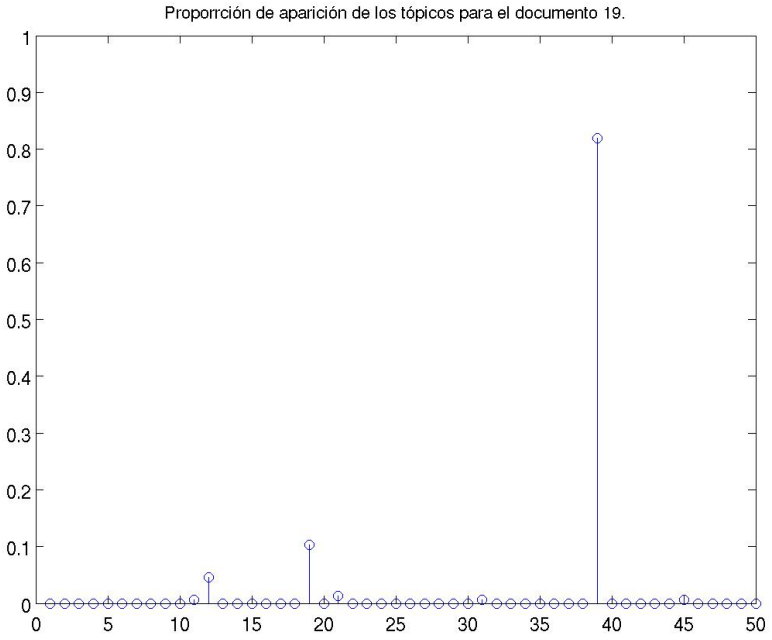
Rank	Distancia	Doc Id.
1	-22.589	55
2	-220.873	63
3	-223.710	69
4	-356.444	93
5	-411.197	24
⋮	⋮	⋮
97	-2054.109	40
98	-2172.625	92
99	-2821.194	31

Cuadro 4.12: Similitud entre el documento 19 y los 99 documentos restantes de la figura 4.10.

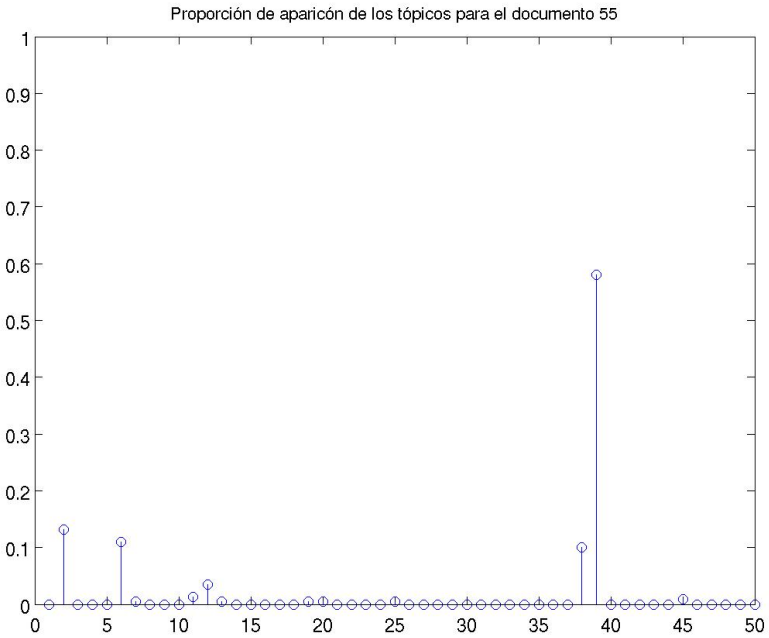
En dicha tabla se advierten los resultados ordenados de mayor a menor de la comparación del nivel de similitud entre el documento 19 y el resto de la muestra.

Las conclusiones obtenidas de esta tabla es que los documentos de la muestra de la tabla 4.12 marcados con los índices $55, 63, 69, 93$ y 24 son los documentos más similares al 19. Por otro lado, los índices $40, 92$ y 31 son los de menor similitud.

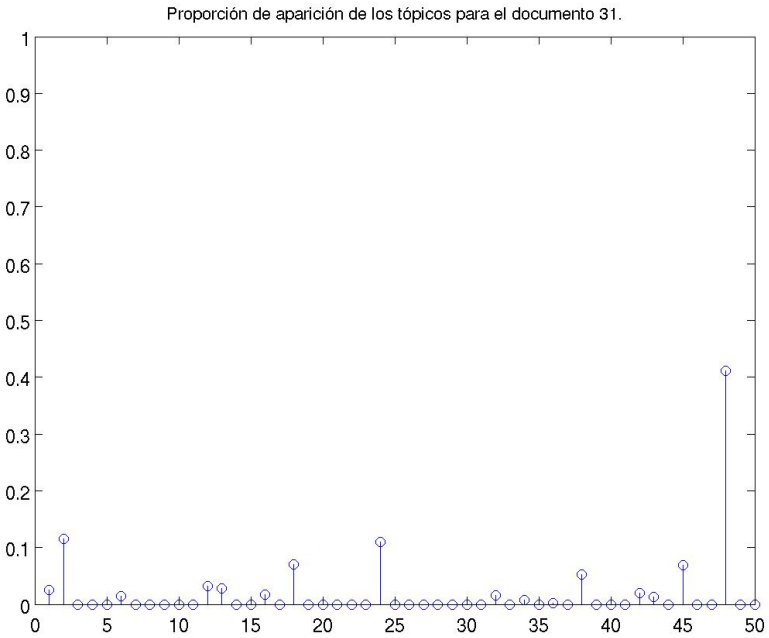
Las figuras 4.14a, 4.14b y 4.14c demuestran la comparación visual entre las distribuciones de probabilidad del documento 19, con respecto al más parecido y el menos parecido.



(a) Documento 19



(b) Documento 55



(c) Documento 31

Figura 4.14: Distribución de los tópicos para los documentos consulta, el más parecido y el menos parecido.

Se puede distinguir que el punto de similitud del documento más parecido y el documento consulta, se encuentra en los valores de las frecuencias dadas por el tópico de la posición 39, El punto de disimilitud entre el documento 19 y el 31 es notable, al observar en la figura 4.14c que en la composición de dicho documento intervienen otros tópicos no considerados en la mezcla del documento 19. Además el tópico 19 tiene poca participación en la mezcla para el documento 31, hecho que matemáticamente permite aumentar la diferencia entre ambos documentos.

Los títulos de los artículos más similares y menos similares según el análisis anterior son: "How are anterior cell migrations guided by mig-13 ?" y "lag-2 is Not Required for the Secondary Cell Fate in Vulval Induction" respectivamente.

Prosiguiendo con este análisis se exponen las figuras 4.15 y 4.16 que muestran partes de los resúmenes etiquetados usando los mismos colores para los tópicos 19 y 39. Se concluye gracias a la figura 4.15 que muchas de las palabras han sido etiquetadas por el tópico 39, mientras que la figura 4.16 no contienen palabras etiquetadas con dicho tópico. En su lugar, de forma esporádica aparecen palabras marcadas con el tópico 45 que en el documento 19 apenas y son seleccionadas.

mig^{19}*-13*^{19} *is a guidance*^{19} *factor*^{19} *that promotes*^{19} *cell*^{19} *migrations*^{19} *in the anterior*^{19} *direction*^{19} *(sym*^{19} *et. al., 1999). Previous work demonstrated mig*^{19}*-13*^{19} *is required*^{19} *for the anterior*^{19} *migrations*^{19} *of the QR*^{19} *descendants*^{19} *and the bdu*^{19} *neurons*^{19} *(sym*^{19} *et. al., 1999). Consistent*^{19} *with the role*^{19} *mig*^{19}*-13*^{19} *in anterior*^{19} *migrations*^{19}*, we hace also found that mig*^{19}*-13*^{19} *also directs*^{19} *the anterior*^{19} *migration*^{19} *of the distal*^{19} *tip*^{19} *cell*^{19} *(DTC*^{19} *in the posterior*^{19} *gonad*^{19} *arm*^{19} *during late L3 .*

Figura 4.15: Parte del resumen perteneciente al documento 55.

Intercellular communication regulates^{45} *the expression cell of the three cell fates (primary, secondary and tertiary) during vulval development. The lin-12 receptor is required for the specification of secondary cell fate, presumably*^{45} *by lateral signalling primary cell (P6.p) and the presumptive secondary cells (P5.p P7.p) . We propose*^{45} *that P6.p expresses lateral signal, yet vulval lateral signal has to be identified.*

Figura 4.16: Parte del abstrac perteneciente al documento 31.

Una aplicación adicional de este modelo, se presenta cuando una vez calculados los valores de las distancias es posible utilizar algún método de aglomeración como

lo es un dendograma. La visualización del dendograma utilizando el promedio de las distancias Kullback Leibler simétricas se aprecia en la figura 4.17. En esta figura se ve que al nivel más bajo que se puede crear hasta 30 grupos para los 100 documentos, las etiquetas asignadas a cada documento se indican en la tabla 4.13.

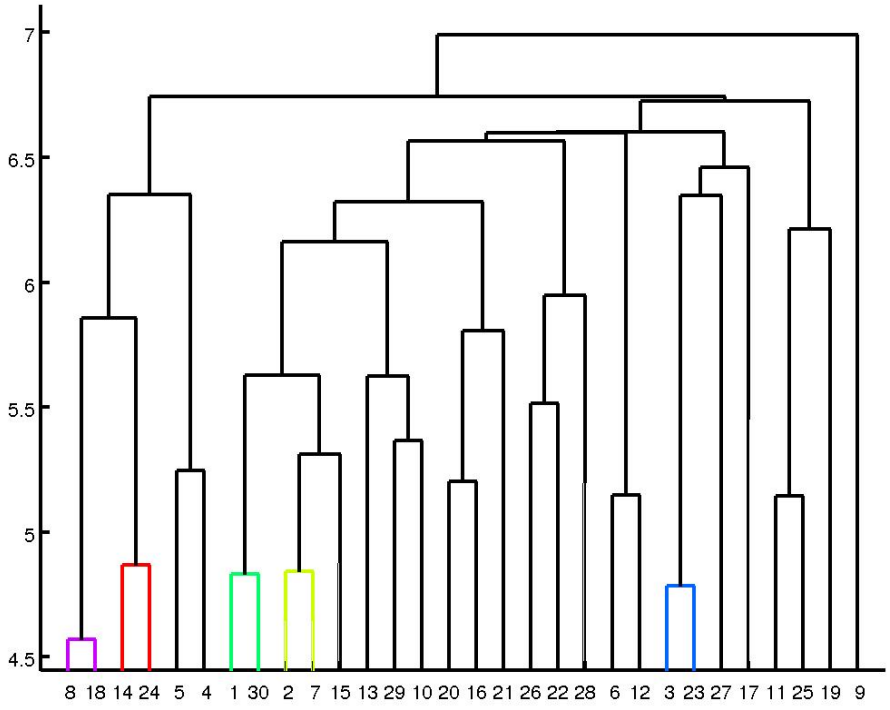


Figura 4.17: Dendograma de la muestra de 100 documentos presentados en la figura 4.10.

Grupo	Num. Docs
1	1,79,85,95
2	2
3	3,4,16,22,80
4	33,78,98
5	5,51,73
6	6,18,23,54,82,90
7	7,59,68,75
8	8,10,63
9	9,12,24,31,64,87,99
10	35,61,100
11	11,28,34,52,53,70,93
12	37,81
13	13,49
14	14,17,69,84
15	15,32,72

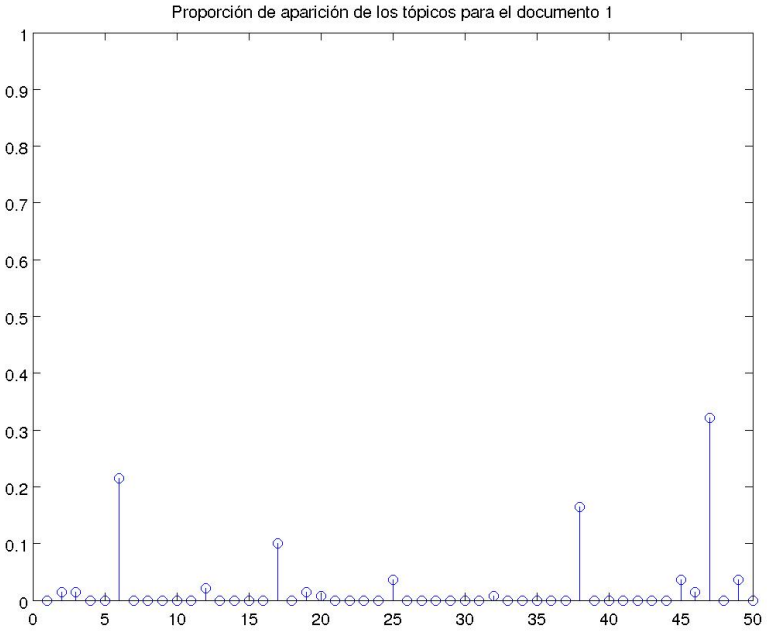
Grupo	Num. Docs
16	38,44,48,67
17	41
18	42,94
19	19,55
20	20,65,77,89
21	21,36,57
22	43,74,76,88
23	46,50
24	71
25	25,40,62
26	26,39,56
27	27,58,86
28	83,92
29	29,60,66,91,96,97
30	30,45,47

Cuadro 4.13: Agrupamiento de los 100 documentos de la figura 4.10.

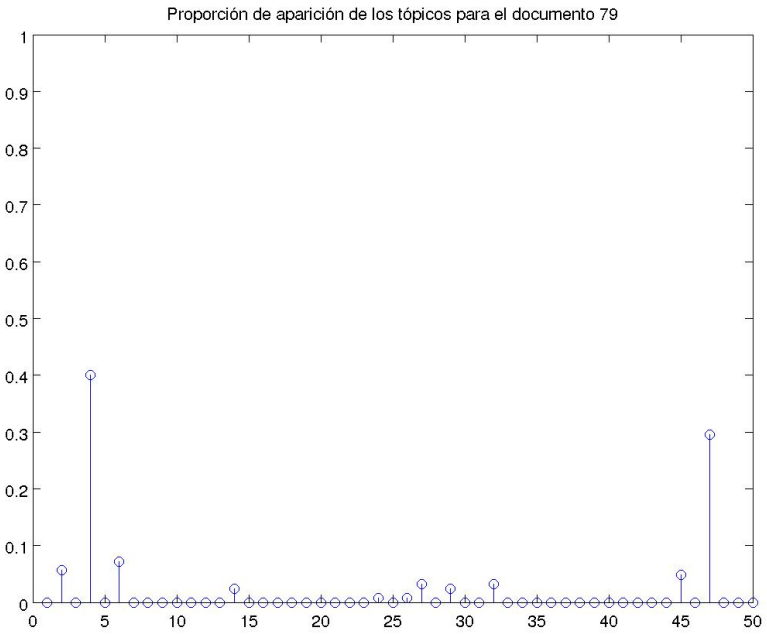
Una forma de validar los resultados de dicho agrupamiento usando como medida de disimilitud la distancia KL, es verificar la estructura que encuentran los dendogramas en las distribuciones de probabilidad de los tópicos para cada documento.

A continuación se muestra un pequeño análisis de cuatro documentos para algunos de los grupos encontrados.

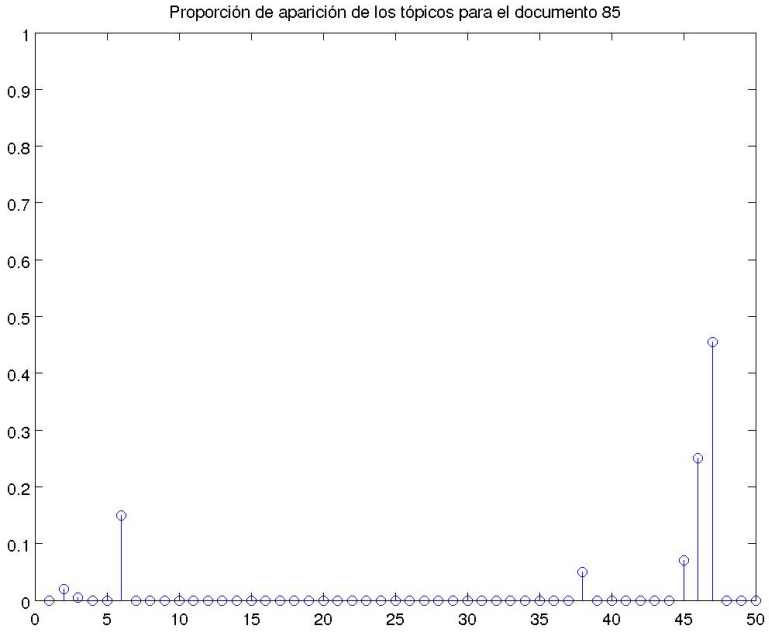
Considere los documentos que conforman el grupo etiquetado como 1, que se observa en la figura 4.18. Se distingue que el patrón a seguir para este conjunto de datos, se encuentra en la existencia de una alta probabilidad de aparición del tópico 6 y el tópico 47. Además, de forma no tan frecuente pero consistente aparece también el tópico 18 para los documentos $1, 85$ y 95. La aparición de estos tópicos no tan importantes producen un efecto de matizado que cambian el significado y contexto de cada documento.



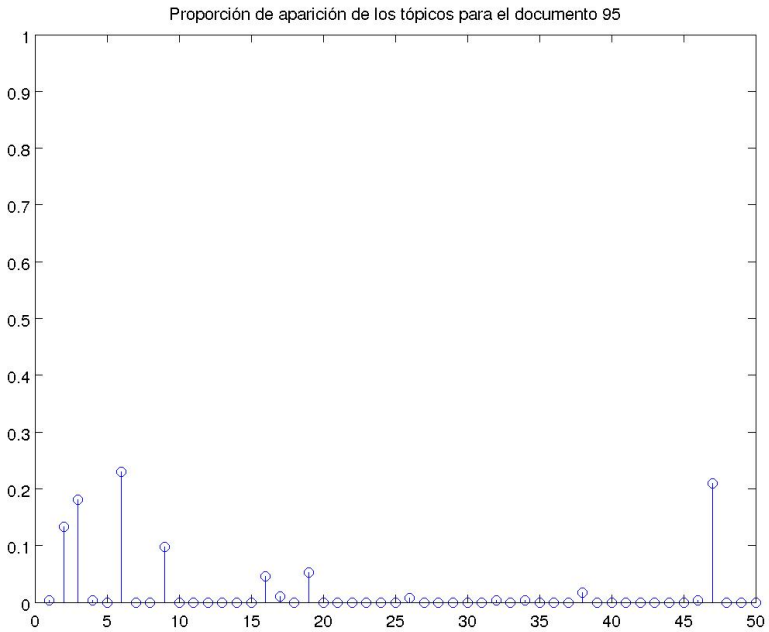
(a) Doc. 1



(b) Doc. 79



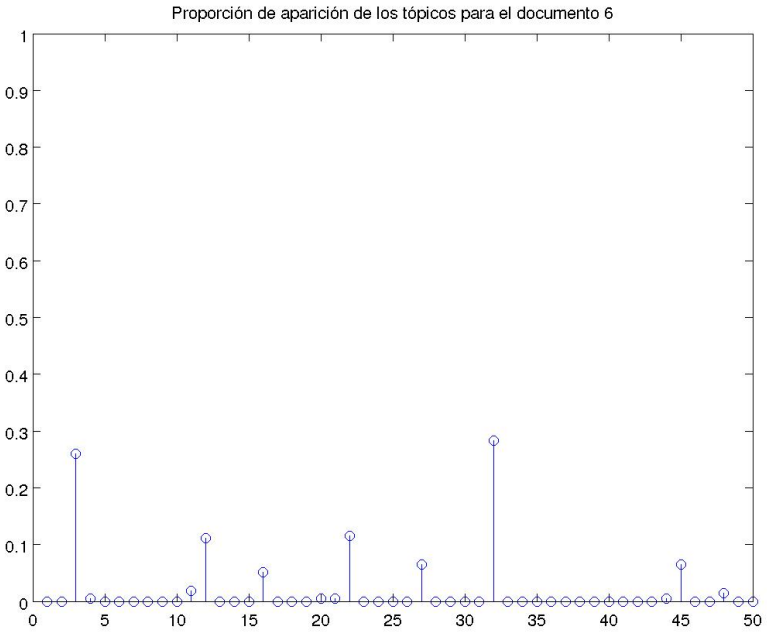
(c) Doc. 85



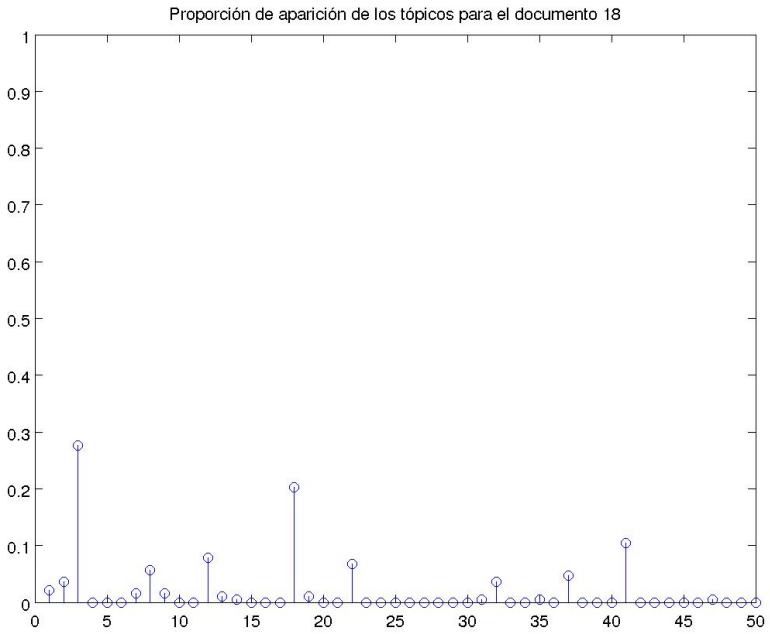
(d) Doc. 95

Figura 4.18: Distribución de los tópicos para los documentos del grupo 1.

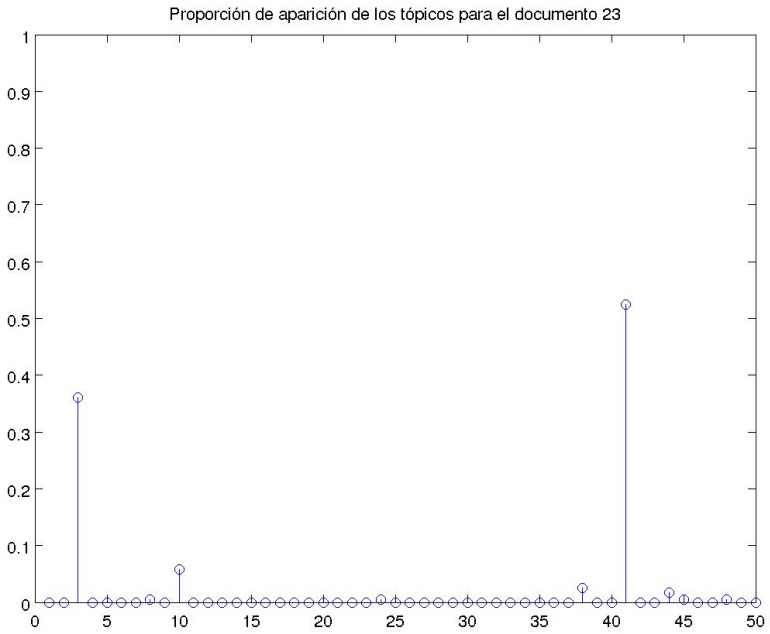
De igual forma la distribución de probabilidad de los documentos del grupo etiquetado como 6, se observa en la figura 4.19. En este grupo aparece de forma consistente el tópico marcado como el 3. Se puede percibir la aparición de este como predominante, en las figuras 4.19b y 4.19d correspondientes a los documentos 18 y 82 respectivamente. Además en las figuras 4.19a y 4.19c de los documentos 6 y 23 dicho tópico es el segundo más predominante con una participación de casi el 30 % y 40 %.



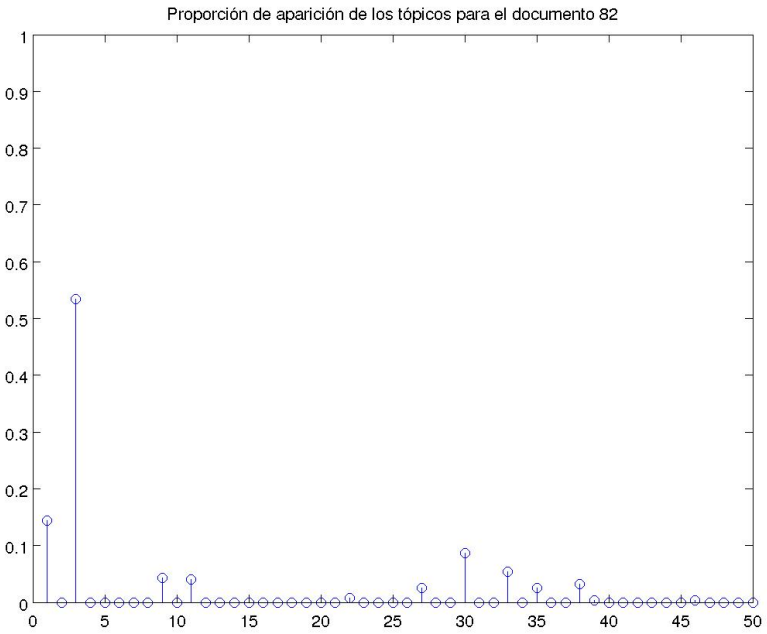
(a) Doc. 6



(b) Doc. 18



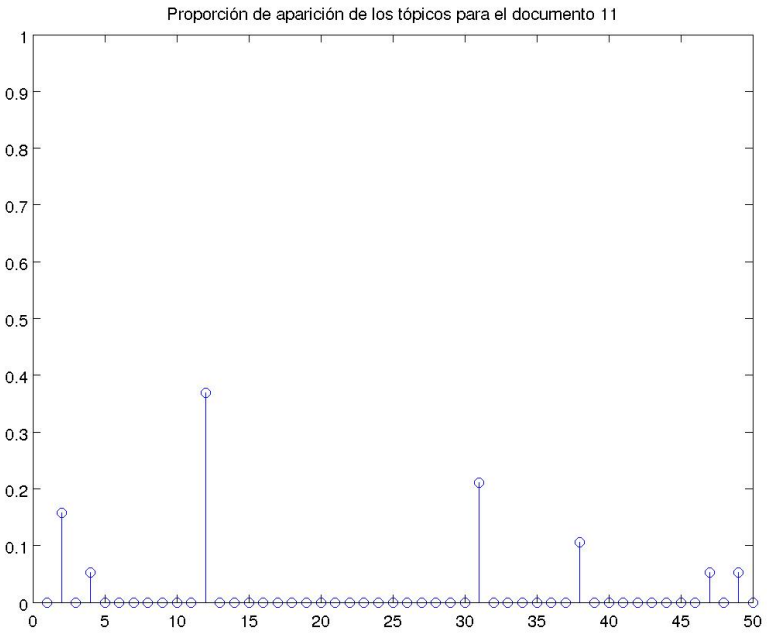
(c) Doc. 23



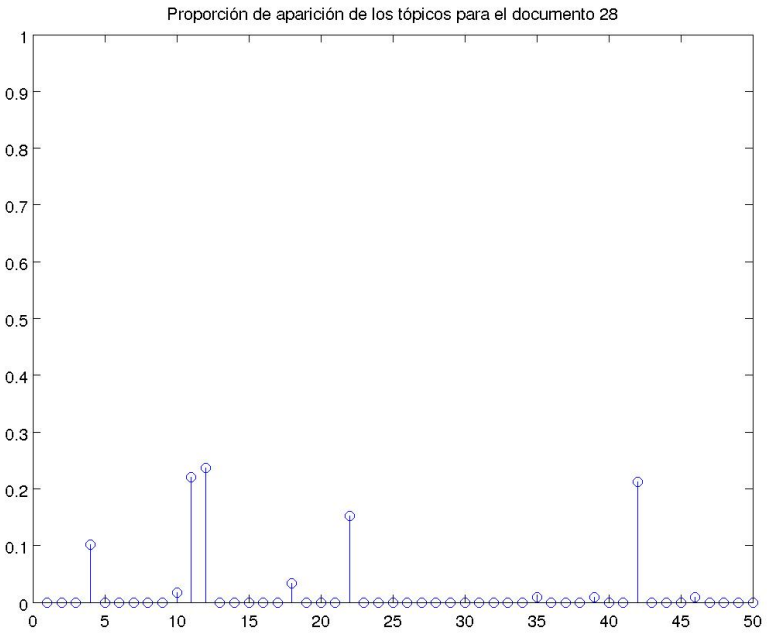
(d) Doc. 82

Figura 4.19: Distribución de los tópicos para los documentos del grupo 6.

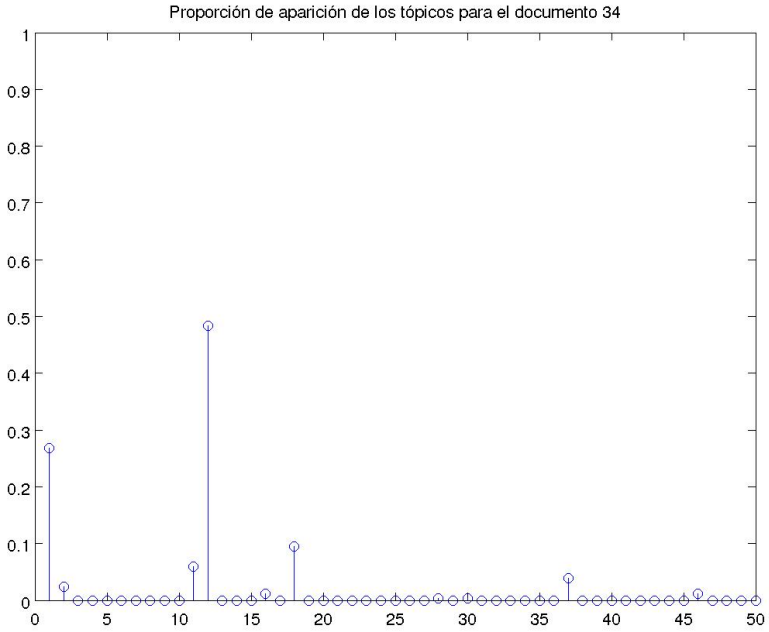
La figura 4.20 muestra las distribuciones para el grupo 11. En este, se nota la aparición frecuente y consistente del tópico 12, además de otros tópicos que ayudan a formar la mezcla. En este caso el tópico 12 ha sido el predominante, aunque no obstante se advierte que para el documento 28, la mezcla se encuentra formada básicamente por 4 tópicos que tienen altas probabilidades de aparición.



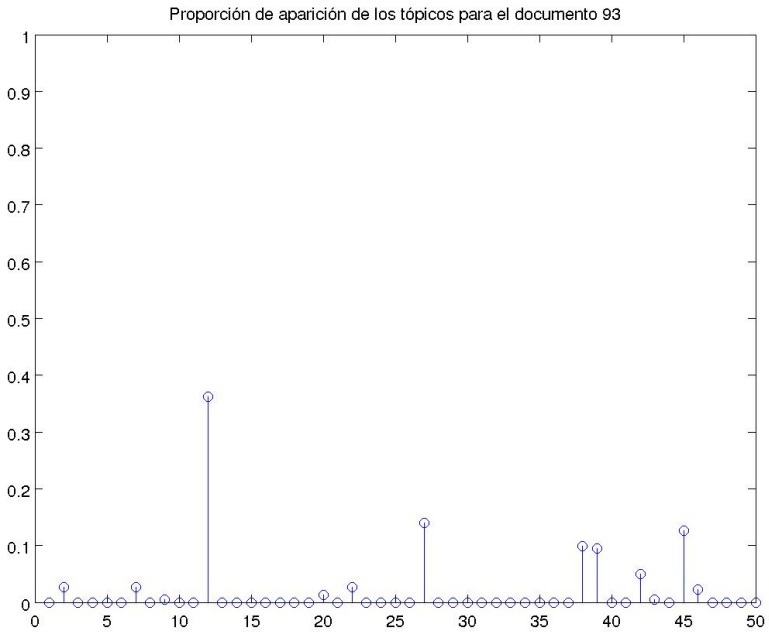
(a) Doc. 11



(b) Doc. 28



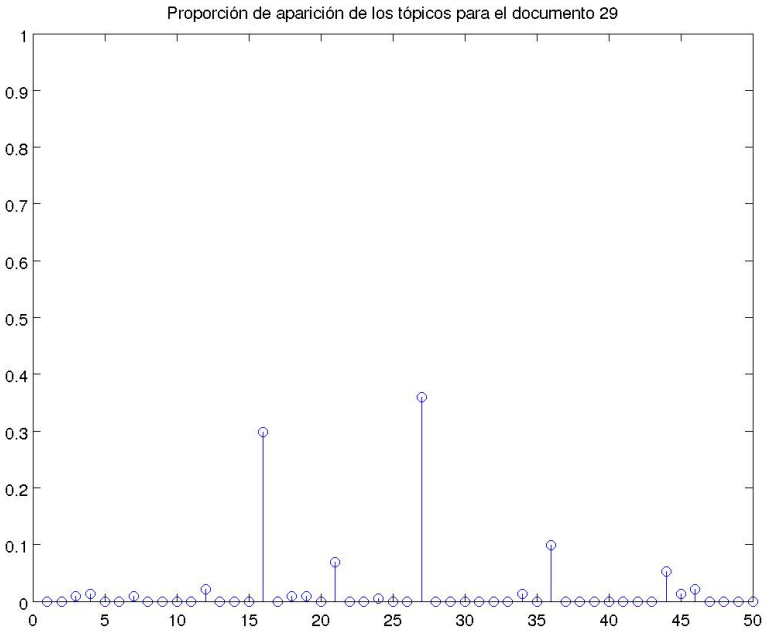
(c) Doc. 34



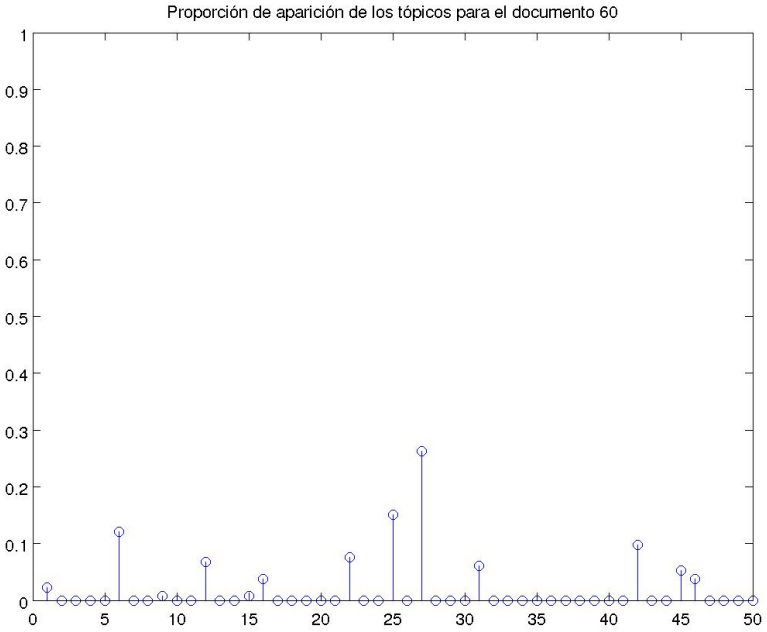
(d) Doc. 93

Figura 4.20: Distribución de los tópicos para los documentos del grupo 11.

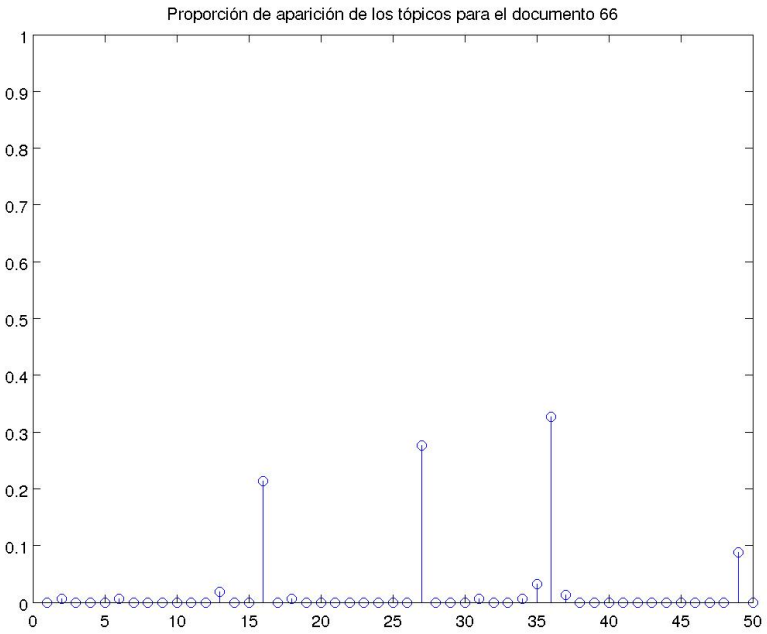
Con respecto al grupo 29, el patrón principal radica en la aparición del tópico 27 como el más consistente. Otros tópicos que también aparecen en menor proporción, conforman a los documentos de este grupo; sin embargo, estos juegan un papel mucho menor que el tópico antes mencionado. Este hecho puede ser verificado en la figura 4.21 en la cual se encuentra la distribución de probabilidad para los documentos 29,60,66 y 91.



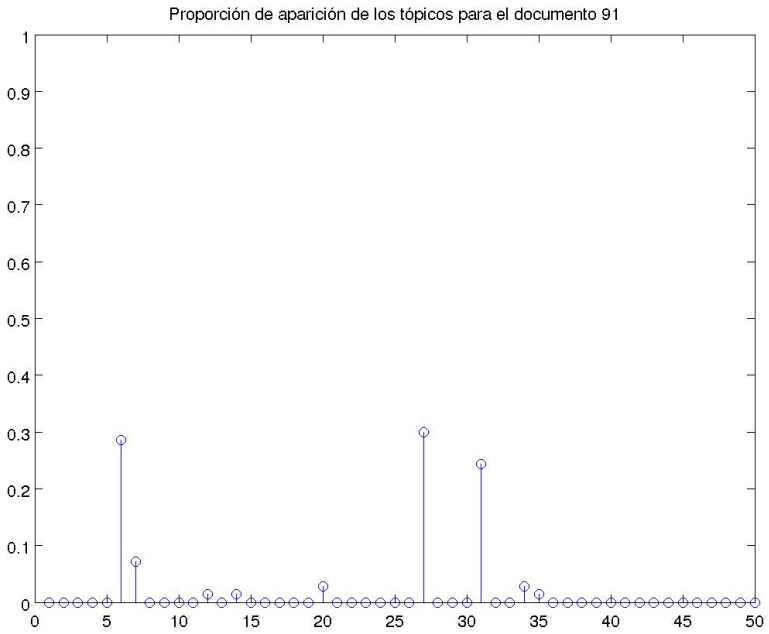
(a) Doc. 29



(b) Doc. 60



(c) Doc. 66



(d) Doc. 91

Figura 4.21: Distribución de los tópicos para los documentos del grupo 29.

Los extractos de los textos de este análisis se incluye en el anexo B.1, donde se puede consultar el tópico al que pertenece cada palabra visualizándo el superíndice con el número asignado correspondiente al tópico.

La última aplicación que se presenta usando el contexto de la muestra de documentos de la figura 4.10, es el de la recuperación automatizada con base a un conjunto de palabras de búsqueda. A diferencia del ejemplo exhibido en la figura 4.14, donde se realiza una comparación entre documentos comparando las distribuciones de probilidades de la mezcla de tópicos que los conforman, esta aplicación se basa en calcular la probabilidad de aparición de las palabras de consulta en los tópicos. Esta es tal vez la aplicación más importante de todas, ya que mediante el uso de un conjunto de palabras de consulta, es posible obtener como resultado un conjunto ordenado de textos considerados como relevantes. Este tipo de estrategias, son similares a las usadas por los buscadores web modernos, que como es bien sabido son capaces de recuperar con suma precisión información cuya temática está íntimamente ligada con la consulta.

Para este fin se eligieron dos palabras con las que usando el procedimiento de la sección 4.5.2 se calculó la probabilidad de la ecuación 4.31. Los términos "asymmetric" y "cell" fueron extraídos del documento 19, y los resultados de la búsqueda se aprecia en la figura 4.12.

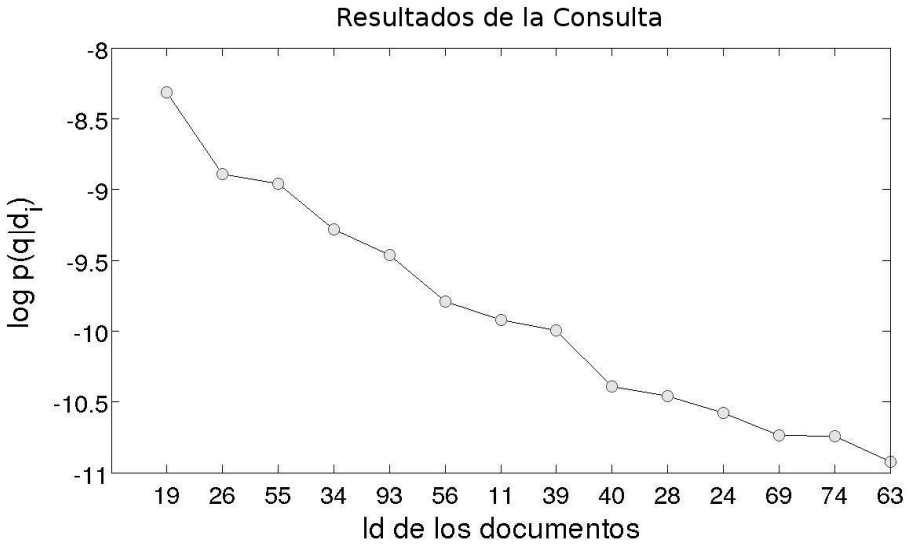


Figura 4.22: Gráfica de $p(q|d_i)$ para la recuperación de documentos.

Debido a que estas palabras fueron tomadas del contexto del documento 19, resulta lógico pensar que dicho documento debe aparecer en los primeros lugares de los valores más altos de la probabilidad.

Por cuestiones numéricas, se ha calculado el $\log p(q|d_i)$ en lugar de calcular directamente $p(q|d_i)$. En la gráfica se señala únicamente los valores de $\log p(q|d_i) > -11$, y en el eje x se anotan los identificadores de los documentos. Es importante hacer notar la aparición del documento 19 en primer lugar, así como, el documento 55 en tercero. Lo anterior, pone de manifiesto la congruencia de los resultados presentados en la tabla 4.12 que sugieren que el documento 55 es el más similar al 19 en términos de tópicos, y que probablemente las palabras de consulta han de provenir de tópicos que comparten en común.

Por último en el apéndice B.2, se puede encontrar los textos de los primeros cinco documentos más relevantes presentados en la figura 4.22.

En síntesis, este capítulo presentó el modelo probabilístico de tópicos conocido como LDA. Posteriormente, se desarrollaron las ecuaciones necesarias para resolver el modelo mediante el uso del algoritmo de muestreo de Gibbs. Una vez resulto el LDA, es cuando el modelo puede ser usado en aplicaciones de minería de datos, a través del uso de herramientas estadísticas. Por este motivo, se presentan varias aplicaciones cuya principal aportación se centra en resumir los datos en un conjunto de documentos que cumplan con ciertas características.

Estas aplicaciones fueron implementadas y usadas para la extracción de información en dos colecciones de datos, cuyo tamaño y complejidad hacen imposible su análisis de forma manual. Entre estas se encuentran:

1. Recuperación de documentos similares.

2. Aglomeración de documentos.

3. Búsqueda y recuperación de información en documentos.

4. Etiquetado automático.

En general, todas estas aplicaciones proporcionan como resultados, información que puede ser interpretada de manera intuitiva por expertos en el área, e incluso en ocasiones por cualquier persona. Además, el encargado de la implementación de estos métodos, no requiere ser un experto en el tema que tratan los datos. Finalmente, mediante las figuras expuestas a lo largo de este capítulo, se verifica que existen diferentes formas de resumir los resultados, para que sea más fácil la interpretación de estos.

Capítulo 5

Modelo de Autores y Tópicos

En este capítulo se explica y desarrolla el modelo generativo de colecciones de documentos conocido con el nombre de "Autores y Tópicos". Este es básicamente una extensión del modelo LDA que fue expuesto en el capítulo 4 y está diseñado para encontrar las distribuciones de probabilidad para las palabras en los tópicos, los tópicos en los documentos y los tópicos para cada autor de forma simultánea. Inicialmente fue dado a conocer en Steyvers *et al.* (2004), Rosen-Zvi *et al.* (2004) y Rosen-zvi *et al.* (2005).

La idea principal de este modelo, es hallar una representación adecuada para inferir los intereses de escritura tomando en consideración una muestra de los documentos. Esta información en general resulta de gran utilidad cuando se intenta responder preguntas tales como:

- ¿Cuáles son los tópicos que más le gusta escribir a ciertos autores?.
- ¿Qué autores tienen afición por los mismos tópicos?.
- ¿Cuáles son los documentos más inusuales para cada autor?.

Con anterioridad se han publicado trabajos que intentan atribuir autoría a un documento, como en Holmes y Forsyth (1995). Sin embargo, en muchas ocasiones solo se intenta trabajar con algunos aspectos más relacionados con el estilo de redacción, es decir, el tipo de palabras usadas, los tipos y frecuencia de Stop Words usadas, longitud de las frases, etc. Además, este tipo de trabajos en general no modela circunstancias, tales como, que un documento sea escrito por varios autores.

Algunas de las aplicaciones de este modelo tienen como objetivo principal el organizar y resumir de forma automática el contenido de grandes colecciones, así como, el estudio de tendencias a lo largo de un intervalo de tiempo.

5.1. Modelo de Gráficas y Proceso Generativo

Al igual que el en modelo LDA, el Modelo de Autores y Tópicos trabaja suponiendo variables aleatorias latentes. En este caso, como se puede apreciar en la figura 5.1, este contiene dos variables latentes representadas por x y z.

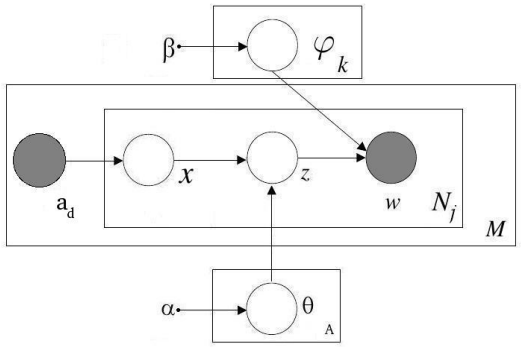


Figura 5.1: Modelo de gráficas probabilísticas del Modelo de Autores y Tópicos.

Introduciendo la notación para explicar el modelo, se define como:

- A es el número total de autores que han intervenido en el corpus.
- W es el número total de palabras diferentes existentes en el corpus.
- M es el número de total de documentos en el corpus.
- N_j es el número de palabras para el j-ésimo documento.
- a_d es la variable aleatoria que representa un vector con los índices de los autores que generan el j-ésimo documento.
- x es la variable aleatoria que señala el índice del autor que escribirá la i-ésima palabra.
- z es la variable aleatoria que elige al tópico de donde provendrá la i-ésima palabra.
- w es la variable aleatoria que define a la i-ésima palabra.
- θ es una matriz de A filas por K tópicos. Cada una de las filas es la distribución de probabilidad de los tópicos para cada autor. Así el elemento (l, k) es la probabilidad de que el l-ésimo autor escriba acerca del k-ésimo tópico.

- α es el parámetro de la distribución Dirichlet a partir del cual se muestrea θ.
- φ es una matriz de K filas por W columnas que representa la distribución de probabilidad de cada palabra para un determinado tópico. Es decir, el elemento (k, w) es la probabilidad de que la w-ésima palabra provenga del k-ésimo tópico.
- β es el parámetro de la distribución Dirichlet a partir del cual se muestrea φ.

Los pasos a seguir para el proceso generativo del Modelo de Autores y Tópicos se enuncia en el algoritmo 5.1.1.

Algoritmo 5.1.1 Modelo de Autores y Tópicos

Muestrear $\theta \sim Dir(\alpha)$
Muestrear $\varphi \sim Dir(\beta)$
para todo Documento en la colección **hacer**
  Escoger los autores que escribirán el j-ésimo documento en a_j.
  para i=1 hasta N_j **hacer**
    Muestrear $x_{j,i} \sim unif(a_j)$
    Muestrear $z_{j,i} \sim Mult(\theta_x)$
    Muestrear $w_{j,i} \sim Mult(\varphi_{z_{j,i}})$
  fin para
fin para

Como se puede apreciar el modelo es muy similar al presentado en el capítulo 3.3, por lo que no es de sorprender que la forma de resolverlo sea en esencia la misma: muestreo de Gibbs.

5.2. Muestreo de Gibbs para el Modelo de Autores y Tópicos

Para esto se escribe la verosimilitud del modelo que está dado según la ecuación 5.1.

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta, A) = \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j} p(x_{j,i}|a_j)p(z_{j,i}|\theta_{x_{j,i}})p(w_{j,i}|\varphi_{z_{j,i}})\right] \left[\prod_{l=1}^{A} p(\theta_l|\alpha)\right]\left[\prod_{k=1}^{K} p(\varphi_k|\beta)\right] \tag{5.1}$$

Esto ocurre,ya que se supone que la variable aleatoria a_j al ser observada es básicamente un parámetro más.

Entonces, al igual que en el modelo LDA se define a las variables aleatorias $x_{j,i}$,$z_{j,i}$ y $w_{j,i}$ como indicadoras. También se emplea el hecho de que $p(\theta_l|\alpha) \sim Dir(\alpha_0)$ y $p(\varphi_k|\beta) \sim Dir(\beta_0)$ y que $p(x_{j,i}|a_j) = \frac{1}{a_j}$. Del mismo modo que en el modelo LDA se reescribe la ecuación 5.1 como en 5.2.

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta, A) &= \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\frac{1}{a_j}\prod_{l=1}^{A}\prod_{k=1}^{K}\theta_{l,k}^{x_{j,i,l}z_{j,i,k}}\prod_{k=1}^{K}\prod_{r=1}^{W}\varphi_{k,w}^{z_{j,i,k}w_{j,i,r}}\right] \\ &\left[\prod_{l=1}^{A}\frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K}\prod_{k=1}^{K}\theta_{l,k}^{\alpha_0-1}\right]\left[\prod_{k=1}^{K}\frac{\Gamma(W\beta_0)}{\Gamma(\beta_0)^W}\prod_{r=1}^{W}\varphi_{k,r}^{\beta_0-1}\right] \\ &= \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\frac{1}{a_j}\right] \\ &\left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\prod_{l=1}^{A}\prod_{k=1}^{K}\theta_{l,k}^{x_{j,i,l}z_{j,i,k}}\right] \\ &\left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\prod_{k=1}^{K}\prod_{r=1}^{W}\varphi_{k,w}^{z_{j,i,k}w_{j,i,r}}\right] \\ &\left[\frac{\Gamma(K\alpha_0)^A}{\Gamma(\alpha_0)^{AK}}\prod_{l=1}^{A}\prod_{k=1}^{K}\theta_{l,k}^{\alpha_0-1}\right]\left[\frac{\Gamma(W\beta_0)^K}{\Gamma(\beta_0)^{KW}}\prod_{k=1}^{K}\prod_{r=1}^{W}\varphi_{k,r}^{\beta_0-1}\right] \end{aligned} \tag{5.2}$$

Por lo tanto, se pueden meter los índices (j, i) hasta los exponentes y denotar $\sum_{j=1}^{M}\sum_{i=1}^{N_j} x_{j,i,l}z_{j,i,k}$ como $n_{(l,k)}$ y $\sum_{j=1}^{M}\sum_{i=1}^{N_j} z_{j,i,k}w_{j,i,r}$ como $n_{(k,r)}$. Además, al reagrupar los productos en la ecuación 5.2 se obtiene 5.3.

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta, A) = \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\frac{1}{a_j}\right]\left[\frac{\Gamma\left(K\alpha_0\right)^A}{\Gamma\left(\alpha_0\right)^{AK}}\frac{\Gamma\left(W\beta_0\right)^K}{\Gamma\left(\beta_0\right)^{KW}}\right] \left[\prod_{l=1}^{A}\prod_{k=1}^{K}\theta_{l,k}^{n_{(l,k)}+\alpha_0-1}\right] \left[\prod_{k=1}^{K}\prod_{r=1}^{W}\varphi_{k,r}^{n_{(k,r)}+\beta_0-1}\right] \tag{5.3}$$

Note ahora, que para eliminar los parámetros θ y φ se integra con respecto a ellos. Debido a las condiciones de independencia condicional de las θ_l que conforman θ, es posible realizar la integral una por una. Lo mismo es aplicable para realizar la integral sobre el parámetro φ. De ahí que la ecuación 5.3 se puede transformar en 5.4.

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w},|\alpha, \beta, A) = \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j}\frac{1}{a_j}\right]\left[\frac{\Gamma\left(K\alpha_0\right)^A}{\Gamma\left(\alpha_0\right)^{AK}}\frac{\Gamma\left(W\beta_0\right)^K}{\Gamma\left(\beta_0\right)^{KW}}\right] \left[\prod_{l=1}^{A}\int_{\theta_l}\prod_{k=1}^{K}\theta_{l,k}^{n_{(l,k)}+\alpha_0-1}\,\delta\theta_l\right] \left[\prod_{k=1}^{K}\int_{\varphi_k}\prod_{r=1}^{W}\varphi_{k,r}^{n_{(k,r)}+\beta_0-1}\delta\varphi_k\right] \tag{5.4}$$

Observe que tanto para θ_l como para φ_k se cumple que:

$$\int_{\theta_l} \prod_{k=1}^{K} \theta_{l,k}^{n_{(l,k)}+\alpha_0-1} \delta\theta_l$$

$$= \int_{\theta_l} \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)} \frac{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}{\prod_{k=1}^{K}\Gamma\left(n_{(l,k)}+\alpha_0\right)} \prod_{k=1}^{K} \theta_{l,k}^{n_{(l,k)}+\alpha_0-1} \delta\theta_l$$

$$= \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)} \int_{\theta_l} \frac{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}{\prod_{k=1}^{K}\Gamma\left(n_{(l,k)}+\alpha_0\right)} \prod_{k=1}^{K} \theta_{l,k}^{n_{(l,k)}+\alpha_0-1} \delta\theta_l$$

$$= \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)} (1) \tag{5.5}$$

Por lo tanto 5.4 se simplifica como 5.6

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, |\alpha, \beta, A) = \left[\prod_{j=1}^{M}\prod_{i=1}^{N_j} \frac{1}{a_j}\right] \left[\frac{\Gamma\left(K\alpha_0\right)^A}{\Gamma\left(\alpha_0\right)^{AK}} \frac{\Gamma\left(W\beta_0\right)^K}{\Gamma\left(\beta_0\right)^{KW}}\right] \left[\prod_{l=1}^{A} \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}\right] \left[\prod_{k=1}^{K} \frac{\prod_{r=1}^{W} \Gamma\left(n_{(k,r)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)}\right] \tag{5.6}$$

Debido a que las variables aleatorias x, z y w solo se encuentran en los términos $n_{,.}$ entonces se puede afirmar lo que se expresa en la ecuación 5.7.

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w},|\alpha, \beta, A) \propto \left[\prod_{l=1}^{A} \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)} + \alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)} + \alpha_0\right)\right)}\right]\left[\prod_{k=1}^{K} \frac{\prod_{r=1}^{W} \Gamma\left(n_{(k,r)} + \beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)} + \beta_0\right)\right)}\right] \tag{5.7}$$

Ahora bien, para usar el muestreo de Gibbs se tiene que encontrar una expresión de la forma $p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1|x_{-(n,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1)$. Usando la definición de probabilidad condicional se sabe que:

$$\begin{aligned} &p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1|x_{-(m,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1) \\ &= \frac{p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1, x_{-(m,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1)}{\sum_{x_{(j,i)}} \sum_{z_{(j,i)}} p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1, x_{-(m,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1} \end{aligned} \tag{5.8}$$

Entonces,

$$\begin{aligned} &p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1|x_{-(m,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1) \\ &\propto p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1, x_{-(m,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1) \\ &= p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w},|\alpha, \beta, A) \end{aligned} \tag{5.9}$$

Es decir,

$$\begin{aligned} &p(x_{(m,n,l_0)} = 1, z_{(m,n,k_0)} = 1|x_{-(n,n,l_0)}, z_{-(m,n,k_0)}, w_{(m,n,r_0)} = 1) \\ &\propto \left[\prod_{l=1}^{A} \frac{\prod_{k=1}^{K} \Gamma\left(n_{(l,k)} + \alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)} + \alpha_0\right)\right)}\right]\left[\prod_{k=1}^{K} \frac{\prod_{r=1}^{W} \Gamma\left(n_{(k,r)} + \beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)} + \beta_0\right)\right)}\right] \end{aligned} \tag{5.10}$$

Pero la parte derecha de esta ecuación puede factorizarse y quedar expresada en función de los valores $x_{(m,n,a_0)} = 1$, $z_{(m,n,k_0)} = 1$ y $w_{(m,n,r_0)} = 1$. Por lo que la

ecuación 5.10 queda como en la expresión 5.11

$$
\begin{aligned}
&p(x_{(j,i,a_0)}=1, z_{(j,i,k_0)}=1 | x_{-(j,i,a_0)}=1, z_{-(j,i,k_0)}=1, w_{(j,i,r_0)}=1) \\
&\propto \left[\prod_{l=1}^{A} \frac{\prod_{k=1}^{K}\Gamma\left(n_{(l,k)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}\right]\left[\prod_{k=1}^{K} \frac{\prod_{r=1}^{W}\Gamma\left(n_{(k,r)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)}\right] \\
&= \left[\prod_{l=1}^{A}\left[\prod_{k=1,k\neq k_0}^{K}\Gamma\left(n_{(l,k)}+\alpha_0\right)\right]\left[\frac{\Gamma\left(n_{l,k_0}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}\right]\right] \\
&\left[\prod_{k=1}^{K}\left[\prod_{r=1,r\neq r_0}^{W}\Gamma\left(n_{(k,r)}+\beta_0\right)\right]\left[\frac{\Gamma\left(n_{k,r_0}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)}\right]\right]
\end{aligned}
\tag{5.11}
$$

Debido a que los términos $\left[\prod_{k=1,k\neq k_0}^{K}\Gamma\left(n_{l,k_0}+\alpha_0\right)\right]$ y $\left[\prod_{r=1,r\neq r_0}^{W}\Gamma\left(n_{k,r_0}+\beta_0\right)\right]$ no dependen de las variables de interés, entonces en general estas son constantes de proporción, por lo que se cumple la ecuación 5.12.

$$
\begin{aligned}
&p(x_{(m,n,a_0)}=1, z_{(m,n,k_0)}=1 | x_{-(m,n,)}, z_{-(m,n)}, w_{(m,n,r_0)}=1) \\
&\propto \left[\prod_{l=1}^{A} \frac{\Gamma\left(n_{l,k_0}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)}\right]\left[\prod_{k=1}^{K} \frac{\Gamma\left(n_{k,r_0}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)}\right]
\end{aligned}
\tag{5.12}
$$

Finalmente, recuerde que $x_{(m,n,l_0)}z_{(m,n,k_0)}=1$ y $z_{(j,i,k_0)}w_{(m,n,l_0)}=1$ por lo que de la ecuación 5.12 se puede escribir de nuevo como en 5.13.

$$
\begin{aligned}
& p(x_{(m,n,l_0)}=1, z_{(m,n,k_0)}=1 | x_{-(m,n)}, z_{-(m,n)}=1, w_{(j,i,r_0)}=1) \\
& \propto \left[\prod_{l=1}^{A} \frac{\Gamma\left(n_{(l,k_0)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)} \right] \left[\prod_{k=1}^{K} \frac{\Gamma\left(n_{(k,r_0)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)} \right] \\
& = \left[\prod_{l=1, l\neq l_0}^{A} \frac{\Gamma\left(n_{(l,k_0)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l,k)}+\alpha_0\right)\right)} \frac{\Gamma\left(n_{(l_0,k_0)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l_0,k)}^{z_{-(m,n)}}+\alpha_0\right)\right)} \right] \\
& \left[\prod_{k=1, k\neq k_0}^{K} \frac{\Gamma\left(n_{(k,r_0)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k,r)}+\beta_0\right)\right)} \frac{\Gamma\left(n_{(k_0,r_0)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k_0,r)}+\beta_0\right)\right)} \right] \\
& \propto \left[\frac{\Gamma\left(n_{(l_0,k_0)}+\alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l_0,k)}+\alpha_0\right)\right)} \right] \left[\frac{\Gamma\left(n_{(k_0,r_0)}+\beta_0\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k_0,r)}+\beta_0\right)\right)} \right] \\
& = \left[\frac{\Gamma\left(n_{(l_0,k_0)}^{z_{-(m,n)}}+\alpha_0+1\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l_0,k_0)}^{z_{-(m,n)}}+\alpha_0\right)+1\right)} \right] \left[\frac{\Gamma\left(n_{(k_0,r_0)}^{z_{-(m,n)}}+\beta_0+1\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k_0,r_0)}^{z_{-(m,n)}}+\beta_0\right)+1\right)} \right]
\end{aligned}
\tag{5.13}
$$

$$p(x_{(j,i,a_0)}=1, z_{(j,i,k_0)}=1 | x_{-(j,i,a_0)}=1, z_{-(j,i,k_0)}=1, w_{(j,i,r_0)}=1)$$

$$\propto \left[\frac{\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)}{\sum_{k=1}^{K}\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)} \frac{\Gamma\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)}{\Gamma\left(\sum_{k=1}^{K}\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)\right)} \right]$$

$$\left[\frac{\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0\right)}{\sum_{r=1}^{W}\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0\right)} \frac{\Gamma\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0 + 1\right)}{\Gamma\left(\sum_{r=1}^{W}\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0\right) + 1\right)} \right]$$

$$\propto \frac{\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)}{\sum_{k=1}^{K}\left(n_{(l_0,k_0)}^{z_{-(m,n)}} + \alpha_0\right)} \frac{\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0\right)}{\sum_{r=1}^{W}\left(n_{(k_0,r_0)}^{z_{-(m,n)}} + \beta_0\right)} \tag{5.14}$$

La ecuación 5.14 representa la distribución condicional con la cual se trabaja mediante el uso del muestreo de Gibbs. De nueva cuenta, se pueden calcular los valores esperados dados los valores de las variables latentes a través de la ecuación 5.15.

$$\begin{aligned} & p(\varphi_k | z_{-(m,n)}, x_{-(m,n)}, w_{-(m,n)}, \beta) \\ & \propto p(w_{-(m,n)} | z_{-(m,n)}, x_{-(m,n)}, \varphi_k, \beta) p(\varphi_k | z_{-(m,n)}, x_{-(m,n)}, \beta) \end{aligned} \tag{5.15}$$

De igual forma, se tiene la expresión $p(\varphi_k | z_{j,i}, x_{j,i}, \beta)$ que puede simplificarse en $p(\varphi_k|\beta)$ debido a que todas variables φ_k llegan a w en caminos head to head desde $z_{j,i}$ y $x_{j,i}$. Además ni w ni ninguno de sus descendientes pertenece al conjunto condicionante entonces $\varphi_k \perp\!\!\!\perp \{z_{j,i}, x_{j,i}\} \mid \emptyset$. Es decir, $p(\varphi_k | z_{j,i}, x_{j,i} | \beta) = p(\varphi_k|\beta)$.

Para la ecuación $p(w_{-(m,n)} | z_{-(m,n)}, x_{-(m,n)}, \varphi_k, \beta)$ se verifica que:

$$
\begin{aligned}
& p(w_{-(m,n)}|z_{-(m,n)}, x_{-(m,n)}, \varphi_k, \beta) \\
&= \frac{p(w_{-(m,n)}, z_{-(m,n)}, x_{-(m,n)}, \varphi_k, \beta)}{p(z_{-(m,n)}, x_{-(m,n)})} \\
&= \frac{p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(z_{-(m,n)}, x_{-(m,n)})p(x_{-(m,n)})}{p(z_{-(m,n)}, x_{-(m,n)})} \\
&= \frac{p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(z_{-(m,n)}, x_{-(m,n)})}{p(z_{-(m,n)}, x_{-(m,n)})} \\
&= p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)
\end{aligned} \tag{5.16}
$$

Ambas relaciones de independencia condicional pueden ser deducidas al ver el modelo de Autores y Tópicos en su versión extendida, presentado en la figura 5.2.

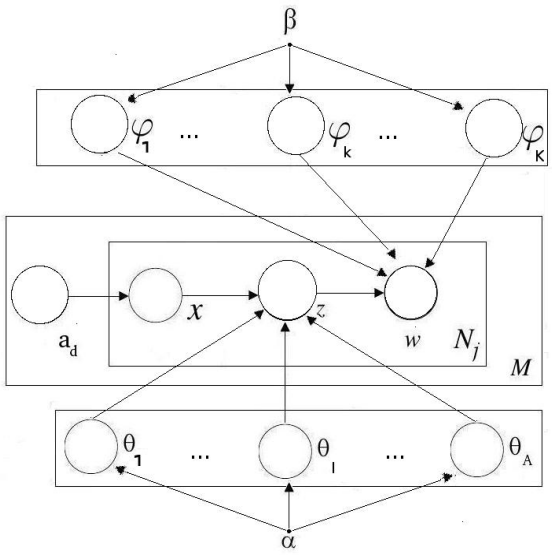


Figura 5.2: Modelo de Autores y Tópicos extendido.

Por lo que la ecuación 5.15 queda como:

$$p(\varphi_k|z_{-(m,n)}, x_{-(m,n)}, w_{-(m,n)}, \beta) \propto p(w_{-(m,n)}|z_{-(m,n)}, \varphi_k, \beta)p(\varphi_k|\beta) \tag{5.17}$$

De igual forma, se sabe que:

$$p(\varphi_k|\beta) = \frac{\Gamma(WB)}{\Gamma(B)^W}\prod_{r=1}^{W}\varphi_{k,w}^{\beta-1} \tag{5.18}$$

También se definió como:

$$p(w_{-(m,n)}|z_{-(m,n)},\varphi_k,\beta) = \prod_{r=1}^{W} \varphi_{k,r}^{n_{(k,v)}^{z_{-(m,n)}}} \tag{5.19}$$

Esto implica:

$$p(\varphi_k|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\beta) \propto \prod_{r=1}^{W} \varphi_{k,r}^{n_{(k,v)}^{z_{-(m,n)}}+\beta-1} \tag{5.20}$$

Debido a esta relación de proporción se puede establecer que:
$p(\varphi_k|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\beta) \sim Dir(\beta')$
donde:
$\beta' = \left(n_{(k,1)}^{z_{-(m,n)}}+\beta, n_{(k,2)}^{z_{-(m,n)}}\beta, ..., n_{(k,W)}^{z_{-(m,n)}}+\beta\right)$.

Si se usa la definición del valor esperado de $\varphi_k|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\beta)$, se encuentra que:

$$\begin{aligned} E\left(\varphi_{k,r}|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\beta\right) &= \int \varphi_k p(\varphi_k|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\beta)\, d\varphi_k \\ &= \frac{\left(n_{(k,r)}^{z_{-(m,n)}}+\beta_0\right)}{\sum_{r=1}^{W}\left(n_{(k,r)}^{z_{-(m,n)}}+\beta\right)} \end{aligned} \tag{5.21}$$

El valor esperado para θ_A es calculando hayando la esperanza de $\theta_A|z_{-(m,n)},x_{-(m,n)},w_{-(m,n)},\alpha$ usando las propiedades de independencia condicional, de forma similar a como se ha hecho en la ecuación 5.16.

5.3. Aplicaciones del Modelo de Autores y Tópicos

Como se ha mencionado anteriormente, el Modelo de Autores y Tópicos es una extensión del modelo LDA. Ambos modelos sugieren la misma estructura, basados en la idea de que las palabras provienen de tópicos; también suponen que los documentos provienen de una mezcla de tópicos.

Estas asunciones ocasionan que ambos modelos tengan que estimar las probabilidades de aparición de las palabras con respecto a los tópicos, generando estructuras similares entre ambos modelos. A diferencia del LDA, el Modelo de Autores y Tópicos tiene la ventaja de inferir distribuciones de probabilidad de los los tópicos para cada autor.

Por este motivo, el Modelo de Autores y Tópicos comparte en su mayoría, las mismas aplicaciones para las que se emplea el modelo LDA y proporciona algunas más. Por ejemplo, es posible realizar el análisis de tópicos por año, recuperación de información, comparación de similitud entre documentos y entre palabras, etiquetado automático de las palabras en el documento, agrupamiento de documentos y análisis de tópicos de moda mencionados en el capítulo 4.5. En adición el Modelo de Autores y Tópicos proporciona un conjunto de aplicaciones como las que se enuncian en capítulos subsecuentes.

5.3.1. Análisis de Tendencia de los Autores por Año

Si la colección de datos cuenta con la información adecuada, es posible hacer un análisis de tendencias de los autores por cada año. Esto se obtiene cuando se grafica la probabilidad de aparición de los tópicos para cada autor divida por años. La utilidad de este análisis radica en poder estimar la evolución de cada autor en sus intereses de escritura. No obstante, hay que recordar que este análisis está basado únicamente en la probabilidad de aparición de los tópicos en los documentos que cada actor ha escrito; por lo que en realidad, esta aplicación simplemente resume la relación que existe entre los tópicos y los documentos escritos por cada autor clasificados por años.

5.3.2. Detección de Documentos poco Comunes

Una vez resuelto el Modelo de Autores y Tópicos es posible usar la matriz θ_A para predecir la aparición de documentos poco comunes, cuando se recibe un documento nuevo, que no se encontraba en la colección.

Este procedimiento se realiza mediante el uso de la perplejidad, mencionada en el capítulo 4.6.2, que sirve para medir el nivel de ajuste que tiene un conjunto de datos con el modelo.

La idea es muy sencilla:

Cada fila de la matriz θ_A representa una distribución de probabilidad multinomial que el modelo encontró para cada autor. Si se evalúa la perplejidad de un nuevo documento contra la distribución de probabilidad para el $l-$ésimo autor, entonces la perplejidad dará una estimación del grado de diferencia entre el documento no observado y los pertenecientes a la colección. Todo esto bajo el supuesto que este nuevo documento fue escrito por el mismo autor. Matemáticamente esto se expresa como:

$$Perplecity(W_d|a) = \exp\left(-\frac{\log p(W_d|a)}{|W_d|}\right) \tag{5.22}$$

Si el nuevo documento efectivamente fue escrito por el mismo autor, entonces en su contenido se espera la aparición del mismo tipo de palabras relacionadas con los tópicos que los autores suelen emplear para escribir. Es por esto, que de la ecuación 5.22 se obtendrán valores pequeños, señalando que el modelo se ajusta adecuadamente a las palabras del nuevo documento. Por otro lado, si el documento contiene palabras que no son comunes en documentos del autor, de la ecuación 5.22 se obtendrán valores muy grandes, señalando que el modelo no se ajusta al documento.

Todo esto se cumple bajo el supuesto de que el modelo estimado por el algoritmo del muestreo de Gibbs para el modelo de Autores y Tópicos, ha convergido adecuadamente, encontrando la verdadera distribución de probabilidad θ_A.

5.3.3. Comparación de Tópicos entre Autores

Si se tienen las distribuciones de probabilidad relacionadas con cada autor, es posible tratar de medir la distancia entre ambas distribuciones trabajando con la distancia simétrica Kullback-Leibler. Si dos autores tienen afición por escribir documentos con tópicos similares, estas distancias serán pequeñas. Por otro lado, distancias grandes demostraran diferencias entre las distribuciones de ambos autores.

De forma matemática se puede expresar esto considerando a los autores $a1$ y $a2$. Las distribuciones multinomiales de probabilidad de ambos autores están dadas por θ_{a1} y θ_{a2}. Entonces, la distancia simétrica Kullback-Leibler se puede escribir como:

$$\begin{aligned} KL(\theta_{a1}, \theta_{a2}) &= \frac{1}{2}\left[D(\theta_{a1}, \theta_{a2}) + D(\theta_{a2}, \theta_{a1})\right] \\ &= \frac{1}{2}\left[\sum_{l=1}^{A} \theta_{a1,l} \log \frac{\theta_{a1,l}}{\theta_{a2,l}} + \sum_{l=1}^{A} \theta_{a2,l} \log \frac{\theta_{a2,l}}{\theta_{a1,l}}\right] \end{aligned} \tag{5.23}$$

Una vez calculadas estas distancias, es posible realizar un ordenamiento y tomar a los autores con mayores distancias y los autores con menor distancia, representando a autores cuyas aficiones por la escritura son diferentes o similares respectivamente.

5.3.4. Etiquetado Automático de Nuevos Documentos para Autores en la Colección

En ocasiones, cuando un nuevo documento llega y los autores de este documento se encuentran en la colección ya analizada, para etiquetar el nuevo documento no es necesario correr todo el proceso del muestre de Gibbs de nuevo.

Debido a que la convergencia de la cadena ya se ha logrado, es posible tomar el estado final de la cadena como estado inicial, y comenzar un muestreo únicamente sobre las palabras de dicho documento. Al transcurrir algunas iteraciones (no tantas como en el primer muestreo), ya se puede obtener un etiquetado. De igual manera, es recomendable realizar este proceso en múltiples ocasiones ya que un solo proceso de muestreo puede generar etiquetas asignadas de forma errónea, debio a la generación de valores aletorios que se realiza como parte del muestreo.

Correr el proceso múltiples ocasiones permite observar un patrón de muestreo, que sugiere los verdaderos valores de las etiquetas de los documentos. En general, muestrear muchas veces mejora la estimación de los valores de las variables latentes, en ocasiones unas 100 iteraciones serán suficientes para poder definir los valores adecuados.

5.4. Experimentos

En este capítulo se presentan algunos experimentos realizados y se comentan los resultados. De nueva cuenta se usaran las bases de datos de NIPS y WormBase que se utilizaron en el capítulo 4.

5.4.1. Experimento con sintético

Esta sección se retoma el ejemplo sintético empleado en el capítulo 4.6.2, donde se generaban imágenes a partir de un patrón de franjas blancas horizontales y verticales. La idea, al igual que en el capítulo antes mencionado, es poder realizar una estimación de los tópicos (los cuales deben asimilarse a las franjas originales), que se muestran en la figura 5.3.

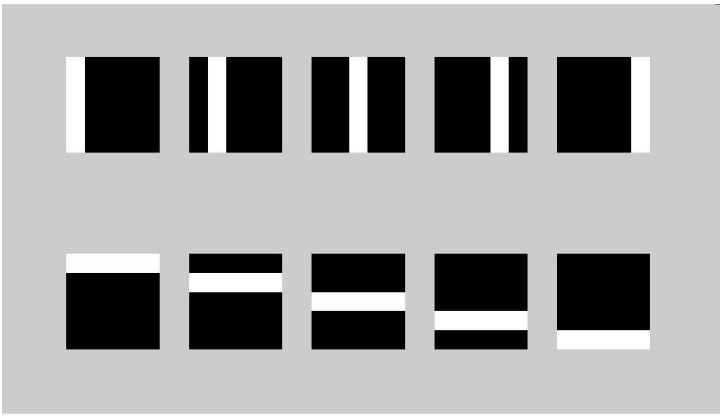


Figura 5.3: Tópicos desde los cuales fueron generadas las imágenes.

Usando los mismos datos y parámetros del experimento del capítulo 4.6.2, se corrió el algoritmo del muestreo de Gibbs para el Modelo de Autores y Tópicos, realizando una estimación de los parámetros. La evolución de los tópicos a través del tiempo del muestreo del muestreo de Gibbs para el Modelo de Autores y Tópicos, se visualiza en la figura 5.4.

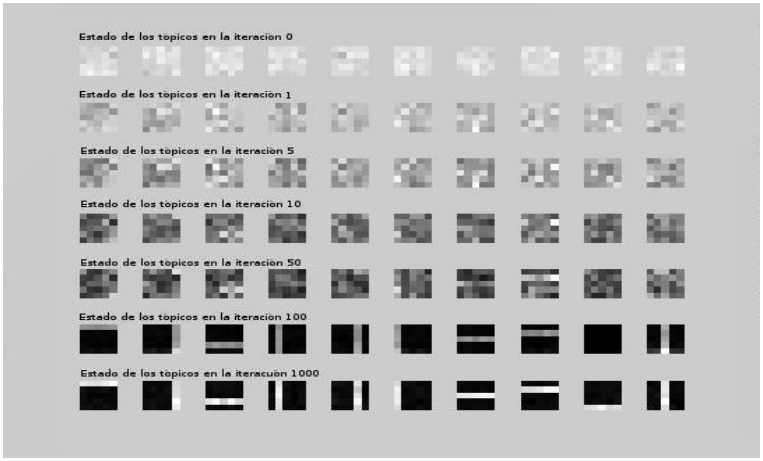


Figura 5.4: Tópicos obtenidos al correr el algoritmo para el modelos de Autores y Tópicos.

Además, se aprecia como de nueva cuenta el algoritmo logra descubrir la distribución buscada cerca de las 100 iteraciones. Subsecuentes iteraciones son usadas para refinar esta aproximación y eventualmente se alcanza un mejor resultado como el que se visualiza en la iteración 1000 de la figura 5.4.

Aunque las líneas horizontales y verticales no se encuentran totalmente definidas en cuanto al color (frecuencia de aparición) al final del proceso, es indiscutible el distinguirlas claramente. La causa de este problema de estimación, radica en el hecho de que ciertos pixeles son encendidos para varios tópicos, ocasionando que el algoritmo haga asignaciones erróneas, otrogando la responsabilidad de dicho píxel a un tópico no responsable. Sin embargo, es notable que esta situación no ocurre con mucha frecuencia, ya que aunque las franjas no son de un color claro sólido, tampoco se visualizan pixeles negros en alguna de las líneas.

Posteriormente, de acuerdo a lo especificado en la ecuación 4.32, se calcula la perplejidad y se extiende la comparación de la figura 4.7 entre el método Variacional Bayesiano y el LDA, incluyendo ahora los resultados de la perplejidad para el Modelo de Autores y Topicos. Esta nueva gráfica se presenta en la figura 5.5.

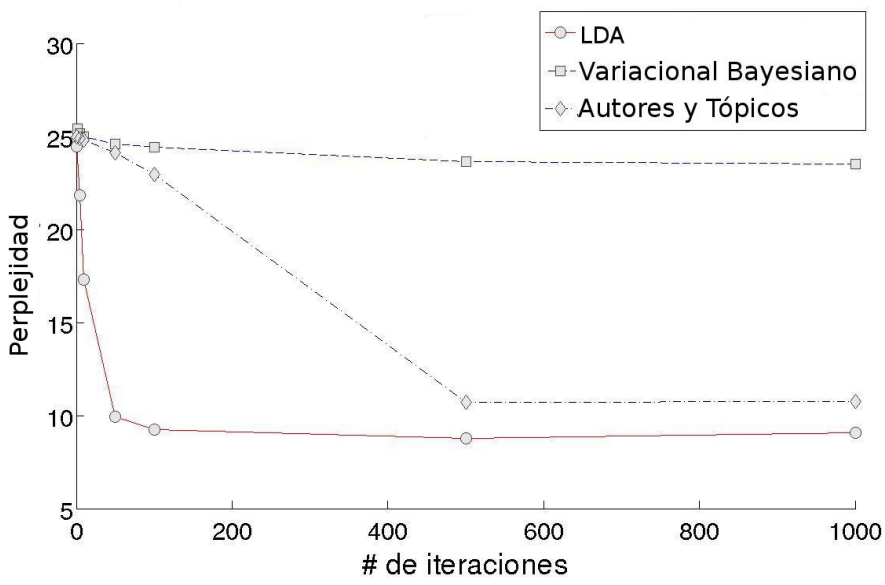


Figura 5.5: Gráfica de la perplejidad aplicada a los resultados obtenidos mediante el muestro de Gibbs para LDA, Variational Bayes y muestreo de Gibbs para el Modelo de Autores y Tópicos.

A pesar de que el LDA es un modelo más sencillo, el Modelo de Autores y Tópicos no representa ninguna mejora en términos de rendimiento, costo computacional o eficiencia de los resultados. Por el contrario, la figura 5.5 pone de manifiesto que los resultados del Modelo de Autores y Tópicos, aunque son mejores que los del algoritmo Variacional Bayesiano, aún no logran superar la calidad de los resultados del algoritmo LDA.

Este comportamiento se atribuye a que en general el modelo de Autores y Tópicos requiere inferir el valor de dos variables latentes (x, z), mientras que el LDA solamente trabaja con una; generando un mayor grado de incertidumbre para los resultados del modelo Autores y Tópicos.

El inferir el valor de un mayor número variables latentes en el proceso de muestreo, causa que el algoritmo de Gibbs requiera una mayor cantidad de iteraciones para lograr la convergencia de la cadena de Markov. Este comportamiento se verifica al observar la evolución de la perplejidad en las iteraciones 1-500, en las cuales la calidad del modelo es inferior a la del LDA y la disminución en el valor de perplejidad es notablemente más lenta.

A pesar de la inferioridad del Modelo de autores y Tópicos en comparación del LDA en cuanto a la estimación de los parámetros relacionados con los tópicos y las palabras, el Modelo de autores y Tópicos es muy usado en muchas colecciones de datos que involucran cuestiones de autoría. Debido que este modelo extrae la misma información que la del modelo LDA y en adición descubre las preferencias sobre los tópicos que los autores tienen al escribir.

Aunque en este ejemplo las cuestiones de autoría no son muy relevantes, es

posible también presentar la distribución de probabilidad para lo autores. Para este caso específico su interpretación sería la responsabilidad que tiene un determinado patrón de barras de contribuir al conteo de dicho píxel. Es decir, cada autor representa a su mismo tópico. Esta distribución de probabilidad se muestra a continuación en la figura 5.6.

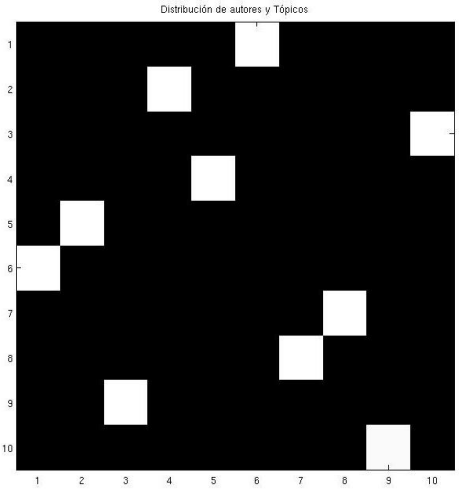


Figura 5.6: Distribución de los autores y los tópicos.

De forma ideal en la figura 5.6 se debería visualizar una linea diagonal descendente de izquierda a derecha. Sin embargo, al ser el algoritmo de muestreo del Gibbs un algoritmo estocástico, difícilmente se podría obtener una estimación de la distribución, que tenga el valor de los tópicos ordenados. A pesar de esto, basta con fijarse que para cada tópico, se tiene una única alta concentración de probabilidad. Es decir, en cada fila solo tiene un único tópico encendido de manera predominante, situación que se puede apreciar claramente en la figura 5.6.

Cabe aclarar que los cuadros negros no significan necesariamente que no se activaron, sino que simplemente tienen poca probabilidad de aparición como se puede apreciar en las tablas 5.1 y 5.2 que muestran las distribuciones de probabilidad para cada uno de los tópicos.

Autor	Tópico	Prob
Autor 1	1	0.0005
	2	0.0002
	3	0.0002
	4	0.0018
	5	0.0024
	6	0.9882
	7	0.0030
	8	0.0011
	9	0.0018
	10	0.0008
Autor 2	1	0.0011
	2	0.0022
	3	0.0008
	4	0.9841
	5	0.0035
	6	0.0014
	7	0.0014
	8	0.0003
	9	0.0050
	10	0.0001

Autor	Tópico	Prob
Autor 3	1	0.0027
	2	0.0004
	3	0.0010
	4	0.0004
	5	0.0002
	6	0.0002
	7	0.0005
	8	0.0016
	9	0.0035
	10	0.9895
Autor 4	1	0.0015
	2	0.0006
	3	0.0039
	4	0.0034
	5	0.9872
	6	0.0010
	7	0.0003
	8	0.0007
	9	0.0012
	10	0.0003

Cuadro 5.1: Distribución de tópicos para los autores de 1-4

Autor	Tópico	Prob	Autor	Tópico	Prob
Autor 5	1	0.0026	**Autor 8**	1	0.0006
	2	0.9903		2	0.0006
	3	0.0005		3	0.0002
	4	0.0014		4	0.0017
	5	0.0013		5	0.0008
	6	0.0001		6	0.0021
	7	0.0002		7	0.9919
	8	0.0011		8	0.0001
	9	0.0012		9	0.0004
	10	0.0013		10	0.0015
Autor 6	1	0.9890	**Autor 9**	1	0.0042
	2	0.0031		2	0.0008
	3	0.0019		3	0.0004
	4	0.0000		4	0.0011
	5	0.0006		5	0.0023
	6	0.0018		6	0.0004
	7	0.0011		7	0.0003
	8	0.0008		8	0.9889
	9	0.0007		9	0.0006
	10	0.0009		10	0.0010
Autor 7	1	0.0042	**Autor 10**	1	0.0042
	2	0.0008		2	0.0008
	3	0.0004		3	0.0004
	4	0.0011		4	0.0011
	5	0.0023		5	0.0023
	6	0.0004		6	0.0004
	7	0.0003		7	0.0003
	8	0.9889		8	0.9889
	9	0.0006		9	0.0006
	10	0.0010		10	0.0010

Cuadro 5.2: Distribución de tópicos para los autores de 5-10

5.4.2. Base de datos de NIPS

Siguiendo el mismo esquema que en el capítulo 4 se han realizado las pruebas con la colección de artículos de la "Neural Information Processing System". Como se menciona en el capítulo 4.6.3.1, esta colección tiene una extensión de 2,301,375 palabras, divididas en 1,740 documentos, escritos por 2,037 autores y un total de 13,649 términos en el vocabulario.

Los resultados mostrados en la figura 5.3 fueron obtenidos mediante el uso del muestreo de Gibbs para el Modelo de autores y Tópicos, usando 50,000 iteraciones y 50 tópicos.

TOP 6	**0.01598**
problem	0.03473
energy	0.02933
optimization	0.02558
solution	0.0252
constraints	0.02268
constraint	0.0198
point	0.01772
function	0.01764
Autores	
Mjolsness_E	0.03853
Rangarajan_A	0.03585
Platt_J	0.03254
Gold_S	0.03038
Barr_A	0.02111
Stolorz_P	0.01635
Derthick_M	0.01286
Leong_H	0.00994

TOP 16	**0.02439**
mixture	0.0309
likelihood	0.02696
data	0.02615
em	0.02305
variables	0.01899
parameters	0.01809
probability	0.01802
log	0.01548
Autores	
Jordan_M	0.05663
Ghahramani_Z	0.05182
Tresp_V	0.03397
Saul_L	0.03375
Frey_B	0.02529
Attias_H	0.02061
Bengio_Y	0.01856
Smyth_P	0.01761

Top 26	**0.01588**
image	0.0559
images	0.04757
face	0.02358
recognition	0.01564
system	0.0126
faces	0.01106
based	0.01098
representation	0.00985
Autores	
Pomerleau_D	0.06921
Baluja_S	0.05283
Spence_C	0.03983
Darrell_T	0.02898
Movellan_J	0.0288
Pentland_A	0.02607
Viola_P	0.02138
Bregler_C	0.01693

TOP 31	0.01996
classification	0.05521
training	0.04794
class	0.03664
classifier	0.03592
data	0.02492
performance	0.02212
classifiers	0.02064
set	0.01977
Autores	
Lippmann_R	0.08779
Chang_E	0.01897
Wettschereck_D	0.01328
Snapp_R	0.01069
Nowlan_S	0.00978
Smyth_P	0.00965
Lee_Y	0.00948
Baram_Y	0.00921

TOP 44	0.01929
gaussian	0.03352
bayesian	0.02727
distribution	0.02639
data	0.02414
prior	0.02288
posterior	0.0214
noise	0.01561
parameters	0.01404
Autores	
Williams_C	0.0789
Bishop_C	0.0528
Barber_D	0.03303
Rasmussen_C	0.02855
MacKay_D	0.02745
Sollich_P	0.0216
Tipping_M	0.01884
Wolpert_D	0.0183

Top 50	0.02444
probability	0.05603
distribution	0.04975
information	0.03277
sample	0.0228
random	0.01958
distributions	0.01288
entropy	0.01268
stochastic	0.01261
Autores	
Tishby_N	0.02189
Kearns_M	0.01239
Watanabe_S	0.01161
Ji_C	0.01108
Shawe-Taylor_J	0.01072
Amari_S	0.01037
Yang_H	0.01009
Zemel_R	0.00866

Cuadro 5.3: Primeras 10 palabras y 10 autores más probables para cada los tópicos .

Como se puede ver, el proceso de muestreo de Gibbs ha logrado encontrar las distribuciones de probabilidad para las palabras, para los tópicos y las de los autores sobre los tópicos.

Una simple inspección visual a las primeras 10 palabras más probables en los tópicos, puede corroborar un buen resultado para la variable φ_k, al notar que la temática es muy notable entre los términos que componen a los tópicos. Por ejemplo, el tópico 6 parece ser un tópico relacionado con el tema de optimización, el tópico 26 trata acerca de imágenes y reconocimiento de rostros a través de imágenes, el tópico 31 tiene por temática el área de clasificación y finalmente el tópico 50 parece estar relacionado con teoría de la información.

Para el caso de la variable θ_A, no es tan simple este análisis. Esto ocurre debido a que un autor puede decidir escribir acerca de un tema específico que está muy ligado a varios tópicos, y a pesar de que los documentos tengan como tema principal un tópico diferente al de mayor probabilidad, es posible que el tópico marcado como más probable sea el que aparece con más frecuencia el documento.

Un ejemplo de este comportamiento se observa en el caso del autor Christopher M. Bishop, cuyos intereses se centran en temas de reconocimiento de patrones. Al ser esta área tan extensa, es posible que aparezca en varios tópicos como ocurre con el tópico 44. Este está relacionado con inferencia Bayesiana, la cual

es una técnica estadística frecuentemente usada en el área de reconocimiento de patrones, por lo que el resultado obtenido en la tabla 5.3 es coherente con la realidad.

Además, el título de los documentos que un autor suele escribir, en ocasiones es de igual manera descriptivo. Los títulos de los primeros 3 autores más probables para cada tópico se presentan en el apéndice C.1. En él se observan los títulos del autor Richard P. Lippmann, donde para las personas familiarizadas con el área de ciencias de la computación, es notable que pertenezcan al área de clasificación.

Debido a que lo más interesante para este modelo radica en sus atributos de autoría, con frecuencia se presentan estos resultados de forma un poco diferente a las mostradas en la tabla 5.3. En el cuadro 5.4 se presentan los cinco tópicos más probables acerca de los cuales suelen escribir los primesos autores de las tablas 5.3.

Autor=Mjolsness_E		
Prob	Top	Palabras
0.29510	6	problem solution optimization energy constraints function
0.17692	49	object objects features recognition feature visual
0.08010	14	image images figure vision pixel pixels
0.07453	10	case work simple general order systems
0.06455	21	figure time shown rate small shows
0.03227	2	data rules clustering cluster rule examples
0.02995	43	network neural networks systems figure results

Autor=Jordan_M		
Prob	Top	Palabras
0.25208	16	mixture data likelihood em variables probability
0.10978	10	case work simple general order systems
0.06222	23	algorithm learning gradient function algorithms error
0.06063	21	figure time shown rate small shows
0.05341	34	performance number results set table test
0.04430	35	learning learn learned task tasks training
0.04401	27	control motor trajectory controller arm feedback

Autor=Pomerleau_D		
Prob	Top	Palabras
0.35357	26	images face image recognition set faces
0.12955	4	units hidden unit network weights training
0.07540	43	network neural networks systems figure results
0.07322	34	performance number results set table test
0.05767	10	case work simple general order systems
0.04310	20	input output layer inputs outputs system
0.03133	21	figure time shown rate small shows

Autor=Lippmann_R		
Prob	Top	Palabras
0.46932	31	classification training class classifier data classifiers
0.17551	34	performance number results set table test
0.12685	21	figure time shown rate small shows
0.03548	47	speech recognition word system training context
0.03358	20	input output layer inputs outputs system
0.02670	18	node nodes tree graph set trees
0.02385	16	mixture data likelihood em variables probability

Autor=Williams_C		
Prob	Top	Palabras
0.36437	44	gaussian bayesian distribution prior data posterior
0.09737	10	case work simple general order systems
0.06047	16	mixture data likelihood em variables probability
0.04639	33	function functions data basis regression linear
0.04545	4	units hidden unit network weights training
0.04139	34	performance number results set table test
0.03951	18	node nodes tree graph set trees

Autor=Tishby_N		
Prob	Top	Palabras
0.22659	50	probability distribution information sample optimal
0.12869	10	case work simple general order systems
0.12569	21	figure time shown rate small shows
0.07118	34	performance number results set table test
0.06178	38	model models data parameters modeling based
0.05002	5	vector vectors code error information coding
0.04831	29	recognition character distance tangent characters digit

Cuadro 5.4: Autores con los tópicos más probables.

Este tipo de representación de la información, resulta más intuitiva para com-

parar intereses de los autores, ya que personas con intereses similares deben tener distribuciones de probabilidad similares. De nuevo el uso de las distancia Kullback-Leibler es una buena herramienta medir la similitud entre autores.

Verbigracia, al hacer una comparación entre la distribución de probabilidad para el autor Jordan_M en esta colección, se obtienen los resultados presentados en la tabla 5.5. La tabla 5.5a presenta a los primeros 10 autores más parecidos, mientras que la tabla 5.5b presenta a los autores ubicados en la posición 990 a la 1000 basados en un ordenamiento de menor a mayor de las distancias Kullback-Leibler.

Distancia KL	Autor
1.0e-05 *0.0234	Williams_C
1.0e-05 *0.064	Doya_K
1.0e-05 *0.1101	Frey_B
1.0e-05 *0.1151	Movellan_J
1.0e-05 *0.1335	Darrell_T
1.0e-05 *0.1739	Kawato_M
1.0e-05 *0.1832	Schraudolph_N
1.0e-05 *0.2537	Linsker_R
1.0e-05 *0.3245	Tenenbaum_J
1.0e-05 *0.5747	Denker_J

(a) Los autores más parecidos

Distancia KL	Autor
0.0526	Kvale_M
0.0527	Miikkulainen_R
0.0529	Cohen_J
0.0529	Fritzke_B
0.0531	Parga_N
0.0532	McCabe_S
0.0538	Zhao_J
0.0543	Kerszberg_M
0.0544	Ryckebusch_S
0.0548	Poggio_T

(b) Los autores menos parecidos

Cuadro 5.5: Distancias Kullback-Leibler para el author Jordan_M.

Antes de analizar los resultados es importante recordar, que esta colección en particular contiene documentos cuyos tópicos se encuentran sumamente correlacionados. Es decir, es muy probable que un autor pueda tener asignaciones de casi todos los tópicos en su distribución, pero un grupo de tópicos en particular serán representativos de las preferencias de dicho autor y este grupo es el de mayor importancia.

En la tabla 5.5a se puede observar que el autor Williams_C es el que tiene un mayor parecido con las distribuciones de probabilidad de los tópicos para el autor Jordan_M. Esto es debido a que entre las áreas de interés generales para ambos autores, se encuentran asuntos relacionados con temas de inteligencia artificial, reconocimiento de patrones y aprendizaje máquina, como se aprecia en las respectivas páginas personales de Michael Jordan ubicada en `http://www.cs.berkeley.edu/~jordan/` y de Chris Williams en `http://homepages.inf.ed.ac.uk/ckiw/mypages/res.html`.

Esta similitud se aprecia en la aparición de lo tópicos 10, 16 y 34 entre los primeros siete tópicos preferidos para estos dos autores.

Por otro lado, analizando los intereses del autor Tomasso Poggio (Poggio_T) en su página personal `http://cbcl.mit.edu/people/poggio/poggio-new.htm`, resulta fácil entender el valor obtenido en la medición de la distancia Kullback-Leibler. Los tres autores tienen afinidad por temas inteligencia artificial, lo que explicaría cierta cercanía con la distribución de tópicos de Chris Williams y Michael Jordan. Sin embargo, la afinidad de Tomasso Poggio por asuntos relativos a las ciencias cognitivas y cerebrales, ocasiona que los documentos de este último autor, sean atacados más desde el punto de vista de modelos bioinspirados, que mediante el uso del técnicas estadísticas y matemáticas.

Los títulos de los documentos de Tomasso Poggio y su lista de tópicos preferidos se enuncian a continuación:

Tomasso Poggio
Learning a Color Algorithm from Examples
A Network for Image Segmentation Using Color
Extensions of a Theory of Networks for Approximation and Learning.
3D Object Recognition: A Model of View-Tuned Neurons
Just One View: Invariances in Inferotemporal Cell Tuning

(a) Títulos en la colección del autor Tomasso Poggio

Autor=Poggio_T		
Prob	Top	Palabras
0.18263	14	image images figure vision pixel pixels
0.15732	10	case work simple general order systems
0.11604	21	figure time shown rate small shows
0.09151	34	performance number results set table test
0.08489	8	visual motion cells field receptive direction
0.05763	49	object objects features recognition feature visual
0.03777	35	learning learn learned task tasks training

(b) Títulos en la colección del autor Tomasso Poggio

Cuadro 5.6: Información del autor Tomasso Poggio.

A pesar de tener tópicos en común, como lo son el 10 y 34, estos tienen valores diferentes, los cuales ocasionan que la distancia sea diferente. No obstante, estas coincidencias son las que ayudan a que la distancia no sea tan grande.

Por otra parte, autores como Cor Van Den Bleek (Van-den-bleek_C), que solo cuenta con un documento titulado "Robust Learning of Chaotic Attractors"; pueden tener una distancia muy grande con los tópicos de interes de Michael Jordan. Esto ocurre debido a este artículo contiene asuntos de redes neuronales, tema muy poco tratado en los documentos de Michael Jordan.

A pesar de que por cuestiones de espacio no se presenta la mezcla completa (con sus 50 componentes); no es posible que un componente de esta distribución sea exactamente cero. Esto es debido a que para poder realizar la estimación con el muestreo de Gibbs, se ha asignado inicialmente una pequeña probabilidad de aparición de los tópicos como se explica en el capítulo 5.2.

Lo señalado anteriormente implica que la distancia entre dos distribuciones difícilmente tomará el valor de ∞, ya que el denominador de la expresión valdrá cuando menos el valor de α_0.

La mezcla de tópicos para el autor Cor Van Den Bleek que ocupa el penúltimo lugar de similitud de acuerdo a los tópicos de Michael Jordan con un valor de la distancia Kullback-Leibler de 3.9556, se muestra en la siguiente tabla:

Autor=van-den-Bleek_C		
Prob	Top	Palabras
0.19935	15	matrix linear vector space component components
0.12745	36	system dynamics time state fixed point
0.11111	39	signal frequency time noise filter auditory
0.08170	13	prediction series data time nonlinear linear
0.08170	18	node nodes tree graph set trees
0.07516	21	figure time shown rate small shows
0.04248	6	problem solution optimization energy constraints function

A continuación se desarrollan más análisis y aplicaciones realizadas a la base de datos WormBase, enfocándose a las aplicaciones relacionadas con los autores, ya que las aplicaciones relacionadas con palabras y tópicos ya han sido discutidas a lo largo del capítulo 4.

5.4.3. Base de datos de WormBase

Siguiendo con el esquema anterior, esta sección exhibe el análisis del Modelo de Autores y Tópicos para la base de datos WormBase. Esta consta de 24,484 documentos escritos por 30,897 autores, conteniendo 4,060,908 palabras y 74,538 términos en el vocabulario, con fechas de publicación que van desde 1910 hasta el 2007.

Una selección de los tópicos arrojados como resultados por el algoritmo de muestreo de Gibbs para el modelo Autores y Tópicos se presenta en la tabla 5.7. Es posible entender la relación entre las palabras en el interior de los tópicos examinados esta tabla.

Como ocurre con el modelo LDA, el Modelo de autores y Tópicos logra descubrir el patrón de aparición entre palabras, ya que a pesar de no ser expertos en el tema de biología de los gusanos, la temática de los tópicos queda clara.

TOPIC 3	0.01465
acid	0.02394
mitochondrial	0.02131
elegans	0.01797
q	0.01644
enzyme	0.01543
c	0.01427
fat	0.01356
mitochondria	0.012
Lemire_BD	0.01177
Clarke_CF	0.00606
Watts_JL	0.00556
Hiroyuki_Arai	0.00504
Kaveh_Ashrafi	0.0046
Pamela_L_Larsen	0.00332
Young_Ki_Paik	0.00308
Vanfleteren_JR	0.00291

TOPIC 18	0.01904
cells	0.06614
cell	0.04977
morphogenesis	0.01779
fusion	0.01741
epithelial	0.01391
hypodermal	0.0138
hypodermis	0.01245
membrane	0.01214
Hall_DH	0.02184
Hardin_JD	0.01401
Hedgecock_EM	0.01401
Podbilewicz_B	0.00993
Jeff_Hardin	0.00967
Benjamin_Podbilewicz	0.00935
Pettitt_J	0.00785
David_H_Hall	0.00721

TOP 29	0.0341
mutations	0.07492
phenotype	0.05166
mutants	0.04032
n	0.03712
alleles	0.03208
mutation	0.02971
mutant	0.02532
allele	0.02149
Horvitz_HR	0.03783
Bob_Horvitz	0.01796
Thomas_JH	0.01312
Wood_WB	0.00872
Riddle_DL	0.00792
Hodgkin_JA	0.00758
Anderson_P	0.00747
Han_M	0.00732

TOP 32	0.0341
data	0.02716
information	0.01561
available	0.014
elegans	0.01346
research	0.01246
new	0.0106
project	0.00968
worm	0.00894
Edgley_ML	0.0121
Durbin_RM	0.00706
Schatz_BR	0.00672
Thierry_Mieg_J	0.00569
Consortium_TCeGS	0.00548
Consortium_Wormbase	0.00494
Theresa_Stiernagle	0.00419
Leilani_M_Miller	0.00403

TOP 48	0.0214
protein	0.14807
proteins	0.10786
domain	0.08554
binding	0.04166
domains	0.03526
terminal	0.02954
interaction	0.02347
amino	0.01737
Ruvkun_GB	0.00327
Zarkower_D	0.00277
Kaech_SM	0.0024
Wadsworth_WG	0.00218
Hiroshi_Qadota	0.00212
Kim_SK	0.00206
Shaw_JE	0.00206
Joohong_Ahnn	0.00194

TOP 50	0.0181
time	0.02017
model	0.01761
using	0.01492
movement	0.012
used	0.01149
analysis	0.01138
individual	0.00907
worm	0.00903
Shuichi_Onami	0.00592
Fire_A	0.00547
White_JG	0.00503
Lockery_SR	0.005
Dusenbery_DB	0.00471
Shawn_Lockery	0.00384
Rex_Kerr	0.00341
Avery_L	0.00333

Cuadro 5.7: Selección tópicos y sus autores para la colección WormBase.

Por ejemplo, el tópico 3 parece hablar de substancias químicas orgánicas, como enzimas y grasa. El tópico 32 habla acerca de investigaciones y datos. El tópico 48, se refiere a asuntos relacionados con proteínas básicamente.

Autor=Lemire_BD		
Prob	Top	Palabras
0.68168	3	acid mitochondrial elegans q enzyme c
0.08015	29	mutations phenotype mutants n alleles mutation
0.04427	33	cycle cell c cdc complex elegans
0.04122	39	human disease protein elegans s function
0.02214	4	e l animals s normal appear
0.02061	1	screen identify identified genes mutants genetic
0.01450	22	worms c temperature l growth s

Autor=Hall_DH		
Prob	Top	Palabras
0.39171	18	cells cell morphogenesis fusion epithelial hypodermal
0.09820	4	e l animals s normal appear
0.09511	14	unc synaptic neurons motor mutants gfp
0.04692	7	muscle unc body wall muscles c
0.04128	31	elegans studies caenorhabditis mechanisms molecular
0.04128	50	time model using movement used analysis
0.04110	22	worms c temperature l growth s

Autor=Horvitz_HR		
Prob	Top	Palabras
0.21994	29	mutations phenotype mutants n alleles mutation
0.19240	4	e l animals s normal appear
0.11193	44	unc gene dpy region e map
0.09057	9	p lin cell vulval cells vulva
0.07653	19	cell ced death cells apoptosis programmed
0.04068	10	cell cells embryos early embryonic stage
0.03253	26	sequence dna kb tc gene cdna

Cuadro 5.8: Primeros autores de los tópico 3,18 y 29 de la tabla 5.7.

Autor=Edgley_ML		
Prob	Top	Palabras
0.51203	32	data information available elegans research new
0.19010	22	worms c temperature l growth s
0.11841	44	unc gene dpy region e map
0.08742	4	e l animals s normal appear
0.02128	50	time model using movement used analysis
0.01018	10	cell cells embryos early embryonic stage
0.00879	26	sequence dna kb tc gene cdna

Autor=Ruvkun_GB		
Prob	Top	Palabras
0.12224	29	mutations phenotype mutants n alleles mutation
0.11819	4	e l animals s normal appear
0.11188	44	unc gene dpy region e map
0.09905	8	expression gfp gene expressed promoter reporter
0.06699	1	screen identify identified genes mutants genetic
0.06650	26	sequence dna kb tc gene cdna
0.05979	10	cell cells embryos early embryonic stage

Autor=Shuichi_Onami		
Prob	Top	Palabras
0.59445	50	time model using movement used analysis
0.27108	2	par spindle embryos cell nuclear p
0.03202	10	cell cells embryos early embryonic stage
0.01174	4	e l animals s normal appear
0.00961	1	screen identify identified genes mutants genetic
0.00961	23	sperm males male oocytes hermaphrodites spe
0.00640	5	rnai rna interference elegans c gene

Cuadro 5.9: Preferencias de los primeros autores de los tópico 32,48 y 50 de la tabla 5.7.

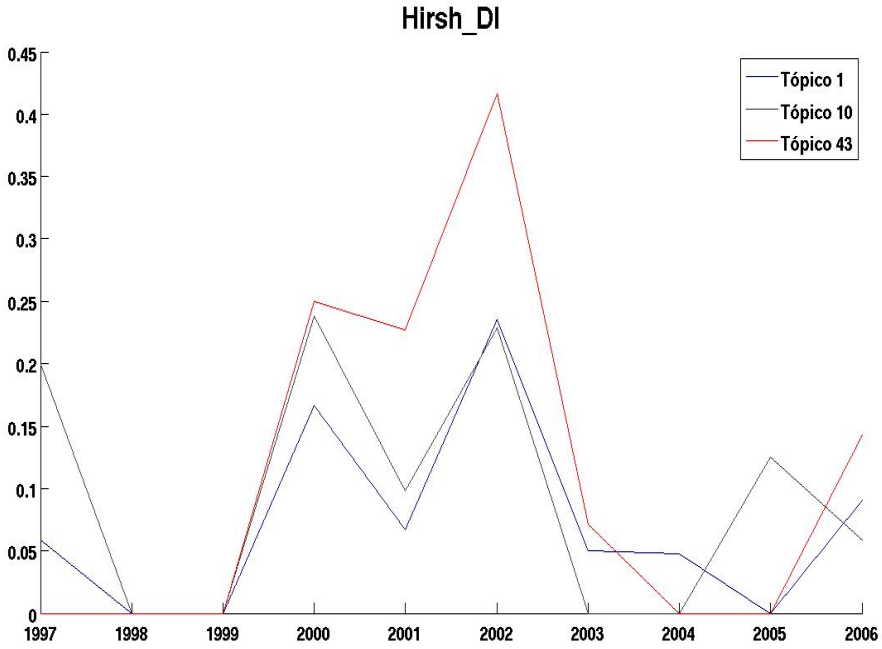
Las tablas 5.8 y 5.9 contienen la información de los tópicos preferidos para los primeros autores de la tabla 5.7.

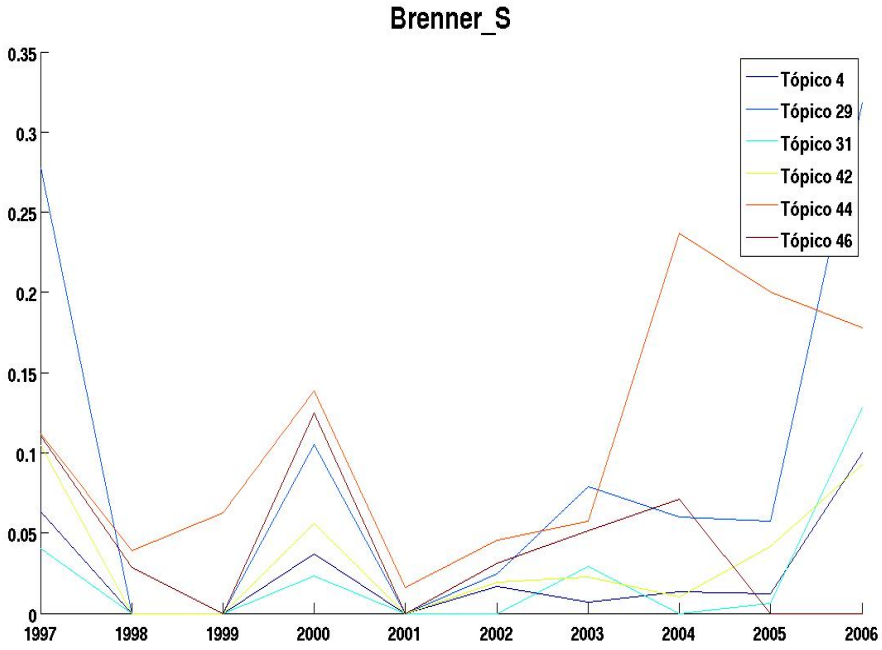
El apéndice C.2 se presentan algunos de los títulos para los primeros tres autores de cada tópico que se señalaron en la figura 5.7. En este caso en particular, los títulos de los documentos no suelen ser tan descriptivos, por lo que esta técnica es usada para catalogar a los autores con base en los temas sobre los que escriben.

Esta característica del Modelo de autores y Tópicos resulta sumamente útil cuando se trabaja en minería de datos, ya que raramente se conoce de forma adecuada la colección a analizar. Es en este punto donde un análisis cuantitativo basado en frecuencias de aparición, suele funcionar mejor que cualquier tipo de

enfoque cualitativo.

Un análisis de evolución de los tópicos por autor entre los años 1997-2006 fue realizado. Los resultados se presentan en la figura 5.7.





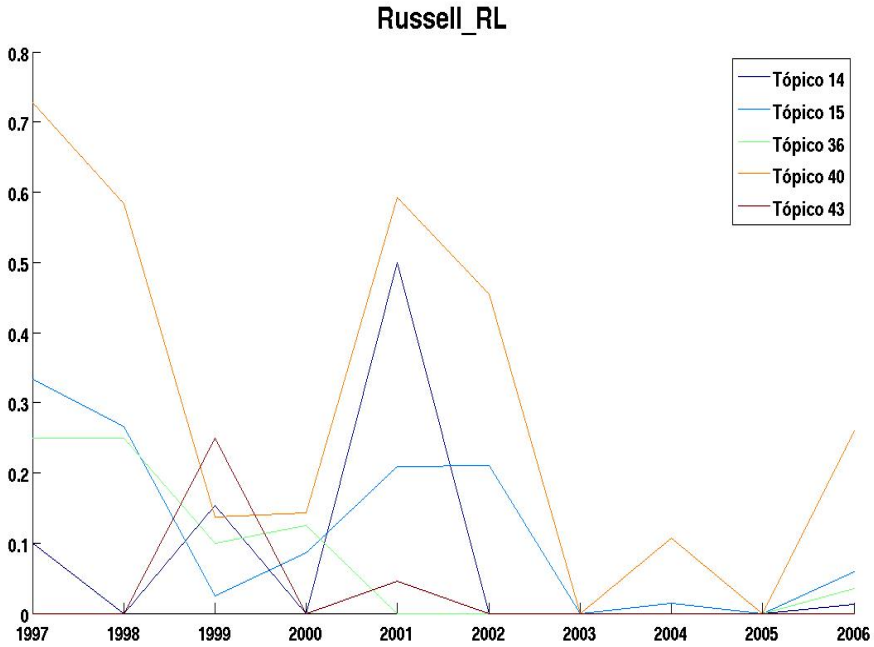


Figura 5.7: Análisis de evolución para tres autores por año.

En las figuras 5.7a,5.7b y 5.7c se grafican las probabilidades de aparición de cada tópico por cada año entre el período indicado anteriormente y para cada autor especificado.Esto es posible hacerlo, ya que se tienen las asignaciones de las variables aleatorias z y x para la colección. De esta forma, se pueden separar las asignaciones por año de publicación, formándose entonces para cada año pequeñas subcolecciones.

Es decir, una vez encontrados los documentos que pertenecen al año en cuestión, se procede a calcula la matriz θ_A contabilizando el número de veces en que cada tópico ha aparecido por responsabilidad de cada autor, dividiéndola sobre el número de veces en las que apareció el tópico en ese año.

Este análisis suele ser muy útil si lo que se busca es tratar de descubrir los patrones de conducta de los escritores o bien analizar los temas que fueron escritos durante un determinado período de tiempo. Además, es posible utilizar un algoritmo de agrupamiento y con ayuda de los expertos etiquetar cada grupo con la finalidad de realizar una clasificación de los autores basados en los temas de los que escriben.

En la figura 5.7 se grafican únicamente los primeros tópicos que aparecen de forma consistente para tres autores de la colección. Se aprecia en las tres gráficas, que a pesar de haber tópicos que ocurren consistentemente en este período de diez años, algunos de estos tienen participación nula. Esta situación de ninguna

forma significa que los autores hayan cambiado de preferencias en lo relativo a su escritura, simplemente indica que durante ese lapso de tiempo los artículos procesados han sido relacionados con otros tópicos, pero que probablemente hacen referencia al mismo tema.

Por otro lado, este análisis tiende a ser más ineficiente en términos de memoria, puesto que se requiere almacenar una matriz θ_A para cada año en el análisis. A pesar de trabajar generalmente con matrices ralas, en ocasiones, debido a la forma de operar del muestreo de Gibbs, se pueden registrar tópicos para autores cuyas asignaciones son muy poco frecuentes, pero que al ser diferentes de cero, requieren ser almacenadas. Para solventar este problema en ocasiones suelen eliminarse estas asignaciones cuando no superen el 0.1 % del total, ya que probablemente estas apariciones se deban a idiosincrasia de la cadena de Markov.

Otra aplicación muy común, suele ser la estimación de las proporciones de escritura de los respectivos autores, es decir, mediante la estimación de la variable aleatoria x del modelo, se etiquetan las palabras otorgando una responsabilidad de aparición de estas para los autores. Entonces contabilizando el número de asignaciones de las palabras para cada autor y dividendo entre el total de palabras se estima la proporción de escritura individual. Las figuras 5.8 y 5.9 muestran segmentos de los resúmenes de dos documentos seleccionados de forma aleatoria de la colección, etiquetados mediante la salida del algoritmo y posteriormente con base a esta misma variable se realizaron los cálculos para las proporciones como se discute más a detalle a continuación.

> *In studying[1] the evolution[2] of nematode[1] feeding[2] behaviors[2], Steciuk[1] et[1].al[2]. showed[1] that differential[1] activity[2] of homologous[1] (and(structurally[1] identical[1]) neurons[1] lead[1] to different[2] feeding[2] behaviors[1] between C[1]. elegans[1] and P[1]. redivivus[1]. In particular[2], the neurons[1] M4[2] and possibly[1] M5[2] are active[1] to excite[2] the pharyngeal[2] terminal[2] bulb[2] muscle[2] in P[2]. redivivus[1], but not C[2]. elegans[2]. A key[1] question[1] is thus: "What are the molecular[1] mechanisms[1] for the differential[2] neuronal[1] activity[1] observed[1] between P[1]. redivivus[1] and C[1]. elegans[2] ?"*

Figura 5.8: Resumen perteneciente el documento titulado "slo-1 modulation of neuronal activity in the pharynx".

The AMP[2]-activated[1] protein[2] kinase[3] (AMPK[3]) plays[1] a key[1] role[1] in the regulation[3] of critical[3] ratios[1] of ATP[2]:ADP[1] and ATP[1]:AMP[3]. AMPK[3] is activated[1] by any stress[1] treatment[1] that interferes[1] with ATP[2] levels[2]. C[2].elegans[1] aak[3]-2[1] mutants[1] that do not have an active[1] AMP[3] activated[1] protein[3] kinase[3], were found to be more sensitive[3] than wild[3] type[1] worms[3] to killing[3] by either starvation[2], high[1] temperature[2] or mitochondrial[1] poisoning[3] (1[3]). Insulin[3]-like[1] signalling[3] mutants[1] also have altered[1] sensitivity[2] to stress[2]. With regards[2] to aging[2] networks[1], it has recently[1] been shown[2] that lifespan[3] extension[3] caused[3] by daf[1]-2[1]/insulin[1]-like[1] signalling[2] mutations[1] is highly[3] dependent[2] on aak[1]-2[3], and that aak[2]-2[2] and daf[1]-16[3] function[1] in parallel[2] responding[2] to overlapping[3] but different[2] inputs[1] (2[3]). We are interested[3] in the stress[1] response[3] aspect[3] of this interrelationship[3] and have exposed[1] aak[3]-2[1] and insulin[2]-like[3] signalling[2] mutants[2] to various[1] stresses[1]. We have chosen[1] stresses[1] known[1] to result[1] in changes[2] in the AMP[3]:ATP[3] ratio[1] such as heat[3] and inhibitors[1] of mitochondrial[3] respiration[1] and stresses[3] such as hyperosmolarity[3] which have unknown[3] effects[3]. We transferred[2] a transgenic[3] firefly[1] luciferase[2] gene[2] to aak[2]-2[2] (ok524[2]), daf[2]-2[1] (e1370[2]) and daf[1]-16[1] (mu86[3]) mutant[1] strains[1] to enable[1] a rapid[2] real[3] time[2] indication[3] of ATP[3] levels[3] in C[2].elegans[2] following[2] the exposure[2] to the above stresses[1] (3[2], see also Lagido[1] et[1] al[1]. accompanying[1] abstract[3]). In addition[3] we determined[3] viability[3] after 24h[2] stress[2] exposure[3] by probing[3] of the worms[2] with a needle[2]. We will discuss[2] the interrelationship[1] of aak[2]-2[2] and insulin[3]-like[2] signalling[2] in the context[1] of the stress[3] response[1]. (1[3]) Apfeld[2] J[3]. et[2] al[3].

Figura 5.9: Resumen perteneciente el documento titulado "Involvement of aak-2 and insulin-like signalling mutations in the cellular stress response as determined by an in vivo ATP sensor C.elegans strain".

La figura 5.8 presenta el resumen del documento titulado "slo-1 modulation of neuronal activity in the pharynx" que tiene como autores Alan_Chiang y Leon_Avery. Las palabras de color azul con superíndice [1] representan las palabras cuya responsabilidad fue atribuida a Alan_Chiang, mientras que las palabras con el color rojo con superíndice [2] corresponden aquellas imputadas a Leon_Avery. En total se contabilizaron 50 palabras (no contando las Stop Words), entre las cuales 30 de ellas fueron relacionadas con Alan_Chiang definiendo una probabilidad de asignación de 0.6. Para Leon_Avery se obtuvieron 20 asignaciones que en proporción representan una probabilidad de 0.4.

Por otro lado, la figura 5.9, muestra el etiquetado automático del resumen que

fue escrito en compañía de 3 autores, cuyos nombres son: Anne_Glover, Cristina_Lagido y Debbie_McLaggan. Las frecuencias de asignación por palabras son de 60,49 y 57 apariciones que dan lugar a probabilidades de 0.3614, 0.2951 0.3433 para cada autor, asignándoles colores verdes, azul y amarillo con los subíndices [1], [2] y [3] respectivamente.

Como se señalo anteriormente, cuando se trabajan colecciones de datos muy grandes, las aplicaciones más importantes son aquellas relacionadas con recuperación automática de información, ya que resuelven los problemas que implica obtener información de manera rápida.

Debido a la dualidad del Modelo de autores y Tópicos en su comportamiento similar al LDA, es posible utilizar la misma técnica descrita en la sección 4.5.2 para la recuperación de documentos. Pero la utilidad principal de modelo Autores y Tópicos radica en la posibilidad de recuperar a los autores más relevantes de la colección a partir de una consulta, conformada por un conjunto de palabras. Esta basa su funcionamiento en el cálculo de la probabilidad $p(A_a|W_q)$, que puede ser obtenido mediante la siguiente ecuación:

$$\begin{aligned} p(A_a|W_q) &= \sum_{i=1}^{N} p(A_a|W_i) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(A_a|z_k)p(z_k|W_i) \end{aligned} \tag{5.24}$$

donde $W_q = \{W_1, W_2 ... W_N\}$ es el conjunto de palabras de consulta, A_a es el a-ésimo autor y z_k representa el k-ésimo tópico.

Mediante la ecuación 5.24 se puede obtener un vector de probabilidades donde cada elemento del vector, asigna una medida de correspondencia de los tópicos sobre los que suele escribir cada autor con las palabras de la consulta.

La tabla 5.10a muestra un ejemplo usando las palabras “phenotype” y “allele”, pertenecientes al tópico 29 de la tabla 5.9. Debido a que ambas palabras se encuentran muy correlacionadas, se supone que el conjunto de autores del resultado debería ser similar a aquel presentado como los autores más relacionados con el tópico 29. Se verifica este hecho comparando los resultados obtenidos en la tabla 5.10a con la sección de los autores para el tópico 29 de la tabla 5.9. Las palabras usadas para el ejemplo 2 de la tabla 5.10b fueron obtenidas del resumen presentado en la figura 5.8. Puede apreciarse que en los resultados de la consulta, el autor Leon_Avery queda situado en el octavo lugar, manteniendo coherencia con la información de los autores contenida en la figura 5.8.

Autor	Prob
Horvitz_HR	0.0780
Bob_Horvitz	0.0338
Thomas_JH	0.0258
Wood_WB	0.0205
Riddle_DL	0.0186
Hodgkin_JA	0.0173
Anderson_P	0.0166
Herman_RK	0.0153
Ruvkun_GB	0.0151
Han_M	0.0151

(a) Resultados de la consulta con las palabras "phenotype" y "allele"

Autor	Prob
Sternberg_PW	0.0486
Avery_L	0.0276
Sommer_RJ	0.0189
Greenwald_IS	0.0166
Kim_SK	0.0143
Fire_A	0.0140
Horvitz_HR	0.0134
Leon_Avery	0.0097
Thomas_JH	0.0091
Ralf_J_Sommer	0.0083

(b) Resultados de la consulta con las palabras "redivivus" y "pharyngeal"

Cuadro 5.10: Ejemplos de recuperación de autores para diversas consultas.

Es importante resaltar que el segundo autor del resumen de la figura quedo situado en la posición 1940. Esto ocurre porque la probabilidad aquí calculada, toma en cuenta la probabilidad de aparición de los tópicos dadas las palabra; sugiriendo que tal vez el autor "Alan_Chiang" no utiliza tópicos que contengan a las palabras de la consulta con probabilidades de aparición altas. Por otro lado, el autor Leon_Avery al parecer está mucho más relacionado con estos tópicos ya que su posición es mejor. Sin embargo, de acuerdo a los resultados existen otros autores, tales como Sternberg_PW, que están más relacionados con los tópicos en cuestión y se convierten en los principales candidatos para los resultados de nuestra búsqueda.

Este tipo de resultados es muy útil cuando se requiere buscar material relativo a un determinado tema. Debido a que los autores obtenidos como resultado de la búsqueda, son aquellos que tienden a escribir más frecuentemente sobre los temas de la consulta, la probabilidad de que el usuario este interesado en su material es alta. Posteriormente, consultar un listado de títulos de los documentos para los autores más relevantes, ayuda a seleccionar el material deseado, reduciendo el tiempo de búsqueda invertido en este proceso.

En resumen, se puede ver que a lo largo de este capítulo, se han desarrollado las ecuaciones pertinentes para poder resolver el Modelo de autores y Tópicos a través del uso del muestreo de Gibbs. Una vez obtenidos los valores de los parámetros, se prosigue a usar el modelo para realizar minería de datos sobre la información de los autores en la colección. Para este fin se expusieron una serie de aplicaciones que de forma estadística logran sintetizar los datos. Usando diversos medios, también es posible presentar interpretaciones de los resultados. En general, estas aplicaciones trabajan mediante el cálculo de distancias y probabibi-

lidades con base a la especificación de alguna consulta y de forma automatizada, reducen la cantidad de datos a un conjunto de documentos considerados como relevantes. La importancia de estas aplicaciones en el contexto de minería, radica precisamente en que proporcionan la posibilidad de reducir el esfuerzo humano para la obtención de información que se encuentra de forma intríseca, más no explícita en la colección. De igual manera, al ser procesos automátizados, los resultados no son suceptibles a errores de tipo humano, que siempre pueden estar presentes cuando se realizan de forma manual.

Capítulo 6

Discución

En los capítulos 4 y 5 se han expuesto, desarrollado y presentado algunos ejemplos del uso y manejo de los modelos LDA y de Autores y Tópicos respectivamente, que permiten resumir la información almacenada mediante el uso de distribuciones de probabilidad.

También se ha demostrado la gran ventaja que representan los modelos probabilísticos para el tratamiento de esta información, ya que mediante el uso de estimaciones de las distribuciones, es posible hacer inferencias sobre diferentes aspectos de la colección. Estas actividades resultan en simplificación de tareas como la indización y búsqueda de información que inicialmente no era accesibles de forma directa.

Las conclusiones que se obtienen de los experimentos realizados, en general, pueden ser clasificadas como:

- Los beneficios relacionados con el método para realizar la estimación de los parámetros del modelo (muestreo de Gibbs), es decir, se discuten temas relacionados con las ventajas que tiene el realizar las estimaciones de las distribuciones de probabilidad mediante el uso del muestreo de Gibbs, por encima de los métodos variacionales que inicialmente se propusieron para dicha tarea.

- Los beneficios de usar un modelo, donde se remarcan las características principales, así como, la forma en la cual la manipulación es usada para ajustar los resultados de acuerdo a las necesidades del análisis.

A continuación se discuten con mayor detalle cada uno de estos aspectos de manera independiente.

6.1. Beneficios del Uso del Muestreo de Gibbs

Para realizar estas estimaciones y resolver los modelos gráficos, se han propuesto una variedad de técnicas estadísticas, que básicamente son divididas en dos familias:

- Métodos variacionales.
- Métodos de muestreo MCMC.

Los métodos variacionales son una familia de métodos para realizar inferencia bayesiana, cuyo interés es hacer una aproximación de la función objetivo, que en este caso está representada por la distribución de probabilidad posteriori de las variables latentes dado el conjunto de datos observados.

Métodos como el EM tienden a resolver un problema de optimización basándose en la evaluación de valores de expectativas. Pero cuando el problema a resolver requiere el manejo de datos en gran dimensionalidad, estos modelos se convierten en un problema imposible de tratar. El algotimo Variacional Bayesiano resuelve el problema de la escalabilidad, sin embargo, plantea otros problemas más complejos a resolver. Esta familia de métodos trabaja siempre con evaluaciones de función, por lo que es necesario proponer un punto inicial donde comenzar a reducir el espacio de búsqueda, llevando al problema de inicialización.

La familia de métodos MCMC, trabaja muestreando de la distribución de probabilidad, por lo que no es neceseario proponer puntos iniciales, librando este tipo de problemas.

Si bien es cierto que el uso de un algoritmo como el muestreo de Gibbs no tiene problemas relacionados con el cálculo de valores iniciales, presenta otro tipo de problemas como el monitoreo de convergencia y el excesivo tiempo de cómputo requerido para obtener los resultados del análisis.

Una gran ventaja de muestreo de Gibbs y en general de los algoritmos MCMC, radica en la precisión de los resultados obtenidos. Ya que las muestras son producidas directamente de la función de densidad de probabilidad, la estimación de la media puede ser tan confiable como se requiera, incrementando el número de muestras para este cálculo. Además, aumentar el tamaño de la población, únicamente afecta el tiempo de procesamiento.

Por otro lado, como se ha discutido en el capítulo 3.4.1, existen una gran diversidad de técnicas usadas para el monitoreo de la convergencia, que en su mayoría son llevadas a cabo de forma concurrente al proceso de muestreo. En muchas ocasiones, las pruebas de convergencia también permiten obtener de forma intrínseca un mejor conocimiento del conjunto de los datos, información extra que puede llegar a determinar mejoras en el ajuste de parámetros que los modelos plantean.

Se debe recordar, que en el contexto de aplicaciones de minería de datos, las colecciones de datos no solo son de gran tamaño; sino que tienen una naturaleza cambiante. El algoritmo de muestro de Gibbs tiene un punto a favor, ya que un resultado obtenido mediante la salida de un proceso de muestreo previo, puede ser usado como un estado inicial para un nuevo proceso de muestreo, aún cuando la colección de datos haya sufrido cambios o aumentos en su tamaño. De hecho, el nuevo proceso de muestreo se espera alcance la convergencia en un período de tiempo menor al anterior (esto dependiendo de la magnitud de los cambios que la colección haya sufrido), haciendo más eficientes los procesos que el algoritmo de Gibbs tendrá que realizar para hacer los cálculos nuevos.

De igual manera, para reducir el tiempo de espera antes de la obtención de resultados, se han propuesto una serie de alternativas tales como las descritas en Newman *et al.* (2006) y Newman *et al.* (2008) que sugieren implementaciones de algoritmos paralelos y distribuidos respectivamente para el modelo LDA. Aunque en general estos algoritmos requieren de mayor poder de cómputo, también proporcionan soluciones a otros problemas. Por ejemplo, un algoritmo distribuido se torna sumamente útil cuando las capacidades de almacenamiento y procesamiento de cualquier equipo de cómputo se ven sobrepasadas. De esta forma, el uso de aglomeraciones de computadoras permite no solo el almacenamiento de los datos, sino también el procesamiento de la información de manera concurrente. Es importante recalcar que aunque estas propuestas han sido desarrolladas pensando únicamente en el modelo LDA, es posible extender las mismas ideas al modelo de Autores y Tópicos sin mayor complejidad, permitiendo que existan versiones paralelas y distribuidas para la búsqueda de autores en la colección.

En resumen, se puede concluir que el uso del algoritmo de muestreo de Gibbs, en principio podría parecer que requiere un excesivo esfuerzo computacional cuando se corre por primera vez. Sin embargo, este tiempo de procesamiento es recuperado en las sucesivas actualizaciones de los resultados, que se originan debido a los cambios que ocurren de manera constante en las colecciones de datos. También se debe recordar que gracias a la aglomeración de computadoras, el tiempo de procesamiento puede ser reducido de manera significativa, permitiendo que el muestro de Gibbs no solo sea una opción escalable, sino que también se eleva la calidad de los resultados, al ser las muestras obtenidas directamente de la distribución de probabilidad que origino los datos observados.

6.2. Beneficios de Usar un Modelo

Cuando un modelo probabilístico es propuesto, siempre existen un conjunto de parámetros que juegan un papel importante en el comportamiento de las dis-

tribuciones de probabilidad. Como se ha explicado anteriormente, los parámetros de los modelos LDA y de Autores y Tópicos entran en la escena cuando se asumen distribuciones Dirichlet a priori para las multinomiales que se manejan en ambos casos. Estos son representados por α y β.

Debido a que los hiperparámetros antes dichos controlan el comportamiento del modelo, es importante predeterminar un valor o conjunto de valores que serán objeto de prueba para la resolución. Se debe recordar, que tanto α como β son vectores de componentes reales. Como se ha explicado en el capítulo 4, por conveniencia se ha propuesto tener todos los valores de las componentes iguales, de esta forma, el número de incógnitas para cada hiperparámetro se ve reducido a simplemente identificar el valor de las componentes y el número de componentes para cada vector. Debido a que el cantidad de entradas de β tiene que ser el mismo que el número de palabras en el diccionario, entonces para cada modelo (LDA y Autores y Tópicos) se tienen 3 parámetros para estimar.

Cuando un determinado conjunto de valores se selecciona para resolver un modelo estadístico, se dice que se ha hecho una selección de modelo. En la práctica, hacer una inversión de esfuerzo para implementar un método de selección de modelo es muy común cuando se trabajan con modelos estadísticos. Sin embargo, en muchas ocasiones esta selección se convierte en un problema inclusive mucho más complejo que el de la estimación original.

En el contexto de minería de datos el problema suele ser inclusive más grave. Debido a que estos métodos de selección de modelo trabajan con funciones de verosimilitud, entonces los resultados siempre estarán basados en mejorar la calidad del ajuste del modelo con los datos. Cabe señalar, que es posible desde el punto de vista humano, que estos resultados no sean considerados como óptimos. Por ejemplo, considere la tabla 4.10, donde se presentan parte de los resultados del análisis de la base de datos de WormBase. Los tópicos están ordenados con cierto grado de granularidad, pero se puede desear tener otro tipo de ordenamiento. Fijándonos en el tópico 50 de la tabla 4.10 que tiene como temática aparente probabilidad bayesiana, en ocasiones se puede desear que este tema sea "dividido" en otros subtópicos. Para realizar este tipo de cambios en el resultado, se debe controlar el número de tópicos del modelo, es decir. el parámetro K. Desde luego que un mayor o menor número de tópicos suele afectar de forma significativa la evaluación de la verosimilitud. Por otro lado, los resultados pueden ser notoriamente mejores desde el punto de vista organizativo o humano. Desde el punto de vista de la minería de datos, el proceso de selección de modelo proporciona una buena idea de valores aproximados a usar, cuando se carece de información acerca de la estructura, aunque por lo general, este solo es útil al comienzo del proceso cuando se realiza una exploración. Con base en los resultados de la etapa exploratoria, con frecuencia se hacen adecuaciones a los parámetros para ajustar los resultados a las necesidades de los usuarios, inclusive frecuentemente se cuenta

con un conjunto de posibles resultados que son usados de formas alternativas dependiendo de la utilidad. En sí, el encontrar el valor adecuado de este parámetro depende de la estructura de los datos a procesar y las necesidades que se traten de cubrir.

Por desgracia, si bien es posible realizar una manipulación del parámetro K para dar un efecto de división de los tópicos, este cambio afecta de manera general a todos los resultados. Ninguno de los dos modelos proporciona algún mecanismo que permita una afectación exclusiva sobre algún tópico específico de manera aislada, sino que, un cambio en el número de tópicos afectará de manera indiscriminada los resultados para todos los tópicos.

El valor utilizado en los componentes de los vectores α y β permiten manipular las probabilidades de aparición de los tópicos y las palabras. Este comportamiento se desprende como consecuencia de la asignación de la distribución Dirichlet como priori de las variables multinomiales y de la asignación del mismo valor a todos los componentes de cada vector. De esta forma, ajustando el valor de la variable α_0 para el vector α y β_0 para β se pueden establecer valores iniciales que favorezcan la aparición de ciertos tópicos o palabras respectivamente, en la parte inicial del muestreo, pero que eventualmente permitan al proceso de Dirichlet encontrar la verdadera distribución. Por lo tanto, los valores de α_0 y β_0 menores a 1 son adecuados.

Este comportamiento es idóneo de forma inicial, si se recurre al hecho de que en colecciones de datos de manera natural, algunas palabras suelen ocurrir con mucha mayor frecuencia que otras, y estas palabras deben aparecer más constantemente que las demás. Un comportamiento similar se exhibe con el manejo de los tópicos. Aunque en general, los valores impuestos a los parámetros suelen jugar un papel importante en los resultados, en la práctica para estos dos modelos el valor inicial no es determinante. El proceso de Dirichlet eventualmente se encarga de llevar al muestreo a la verdadera distribución, entonces los valores iniciales serán empleados únicamente como parámetro de suavizado como se ha explicado en el capítulo 4. Siempre y cuando los valores asignados sean pequeños, los resultados no se verán afectados en demasía.

En síntesis, se afirma que las herramientas estadísticas ayudan en labores de minería de datos, pero aún se tiene que lidiar con el ajuste de los parámetros. El encontrar los valores adecuados para estos en los modelos del LDA y de Autores y Tópicos no es un problema complejo en este contexto, debido a finalmente se realizan constantes cambios que buscan encontrar los mejores resultados que se ajusten más a nuestras necesidades. Si bien el proceso implica esfuerzo en términos de procesamiento, es claro que los resultados pueden justificarlo, al final de cuentas lo más importante es tener la posibilidad de realizar dichas adecuaciones.

Un comentario final podría conjuntar las observaciones hechas a lo largo de este capítulo. Y es que los modelos de gráficas como el LDA y el de Autores y

Tópicos, proporcionan un marco de trabajo para aplicaciones de minería de datos que no solo obtienen resultados confiables, sino que permite realizar una serie de exploraciones, para estudiar diversos aspectos de los datos con los que se trabaja para poder tener un mejor conocimiento de la colección. Además, también se han encontrado formas alternativas a las tradicionales, para resolver estos modelos, que sean más precisos y que se ajusten de manera simple para explotar el poder de cómputo que en estos días ha incrementado.

Por último, se debe resaltar que el uso de estos métodos, permite un conjunto de aplicaciones que facilitan las tareas de recuperación, indexado y clasificación de grandes volúmenes de datos de forma automatizada, reduciendo los recursos humanos de expertos para mantenerlos concentrados en tareas más importantes desde el punto de vista de la interpretación de la información.

Capítulo 7

Conclusiones

Después de una revisión exahustiva de los modelos LDA y de Autores y Tópicos se sugieren las siguientes conclusiones:

El uso de modelos matemáticos y en el caso especial de los modelos estadísticos, facilitan la comprención de forma detallada del proceso que dio origen a los datos. Sin embargo, sugerir un modelo de tópicos para un problema específico, solo es posible si el conocimiento de la colección permite una adecuación de los datos a un proceso que dé un sentido estructural a lo que un tópico representa. Además, estos suelen servir como un proceso de puente, a través del cual los datos observados son analizados usando herramientas estadísticas.

La principal problemática del modelado estadístico radica en la forma de resolver el modelo propuesto. Para afrontar estas situaciones se han planteado una serie de soluciones alternativas entre las cuales se encuentra el muestreo de Gibbs perteneciente a la familia MCMC. A pesar de requerir una gran cantidad de recursos computacionales, existen variantes de estos, que usan algoritmos paralelos y distribuidos para reducir de forma significativa el tiempo requerido para la obtención de resultados finales. A pesar de esto, el esfuerzo computacional sigue siendo considerable, aunque parte de este puede ser recompensado en sucesivas actualizaciones de la información requerida.

La conclusión más importante está relacionada con las aplicaciones presentadas a los largo de está tésis. Y es que todas las aplicaciones utilizadas para ejemplificar el uso de los modelos LDA y de Autores y Tópicos, se basan en propiedades del área de la probabilidad y estadística únicamente. Y es que a pesar del gran tamaño de las colecciones examinadas, es posible resumir la información mediante el uso de las distribuciones de probabilidad planteadas por los modelos.

En general, se puede afirmar que las técnicas de minería de textos aquí men-

cionadas, son simplemente aplicaciones probabilísticas y estadísticas que pueden ser implementadas gracias al uso de distribuciones de probabilidad.

7.1. Trabajo a futuro

En este documento se presento una propuesta para trabajar la selección de modelo que permite estimar valores adecuados para los hiperparámetros del LDA y de Autores y Tópicos. Una segunda propuesta conocida como "métodos no paramétricos" se ha mencionado también en el desarrollo de este trabajo. Como parte de un análisis posterior, sería deseable examinar las ventajas del uso de este tipo de algoritmos, así como, los potenciales beneficios en aplicaciones relacionadas con la minería de datos.

Los modelos LDA y de Autores y Tópicos fueron concebidos originalmente en el contexto de minería de texto, debido a que esta área tiene definido de forma intuitiva conceptos tales como tópico, documento y palabras entre otros. Sin embargo, este tipo de modelos se han aplicado a una gran diversidad de problemas como lo son datos de biología, datos en colecciones de imágenes y objetos 3D, los compuestos químicos entre otros. Desarrollar aplicaciones de recuperación de información y minería de datos para grandes colecciones de este tipo, que además sean implementadas de forma distribuida o paralela.

Por otro lado, a pesar de que las ideas para la implementación de algoritmos paralelos y distribuidos en el modelo LDA no son tan complejas, hasta el día de hoy no se ha presentado alguna propuesta de este tipo para el modelo de Autores y Tópicos. Una buena posibilidad de trabajo futuro radica en realizar una extensión de este modelo al uso de aglomeraciones de equipos de cómputo y demás tecnologías que trabajan de forma concurrente o colaborativa.

Por su parte , la investigación acerca de los modelos de gráficas con variables latentes del estilo LDA y de Autores y Tópicos se han extendido mucho, dando origen a nuevos modelos, que integran capacidades de modelado para diferentes circunstancias y necesidades. Por esta razón, un punto pendiente por atender, se centra en el estudio, implementación y experimentación de diversos modelos de gráficas de variables latentes, para tener una variedad de opciones que permitan atacar diversos tipos de problemas.

Apéndice A

Métodos Variacionales

Cuando se trabaja con modelos probabilísticos, con frecuencia se requiere evaluar la distribución posterior $p(z|x)$ de las variables aleatorias latentes z con respecto a los datos observados x y las evaluaciones de las expectativas calculadas para esta distribución.

Para muchos modelos de interés, no es posible el evaluar la distribución posterior y por lo tanto tampoco se pueden calcular expectativas con respecto a esta distribución. Esto ocurre ya sea debido a que la dimensión de las variables latentes es muy alta y no permite ser manejada directamente o bien, a que esta distribución tiene una forma muy compleja, ocasionando que las expectativas no tengan una forma analíticamente tratable. En el caso de variables continuas, las integrales necesarias pueden ser muy difíciles de calcular de forma analítica o la dimensionalidad puede ocasionar que la integración numérica sea casi prohibitiva. Para el caso de las variables discretas, el maginalizar sobre un conjunto grande de configuraciones para las variables latentes, hace el proceso muy costoso computacionalmente. En estas situaciones, se requiere encontrar otros esquemas que permitan sobrellevar estos problemas. Básicamente las propuestas son dividias en enfoques estocásticas como el MCMC o bien determinísticos como los llamados Métodos Variacionales.

Estos métodos, tienen sus orígenes en el siglo XVIII con el trabajo de Euler, Lagrange y otras personas que trabajaron con cálculos de variaciones. En el cálculo normal, se trabaja encontrando derivadas de funciones. Pensando una función como un mapeo que toma un valor de una variable como entrada y devuelve el valor de la función como salida. La derivada de una función describe entonces cómo la salida varía cuando se realizan cambios infinitesimales a los valores de entrada. Así, se define un funcional como el mapeo que toma una función de

entrada y devuelve el valor del funcional como salida. Un ejemplo de funcional puede ser la entropía $H[p]$ presentada en la ecuación A.1, la cual recibe una función de probabilidad $p(x)$ y devuelve una cantidad. La derivada de un funcional entonces representa la forma en la que el valor de un funcional varia con respecto a cambios infinitesimales en la función Feynman *et al.* (1964).

$$H[p(x)] = \int p(x) \log p(x) \; dx \tag{A.1}$$

Muchos problemas pueden ser expresados en términos de un problema de optimización en el cual se optimiza un funcional. La solución se obtiene explorando el conjunto de soluciones que pueden ser tomadas como entradas del funcional y hallando aquel que lo maximiza o minimiza. Esto se logra restringiendo el rango de funciones que son consideradas como candidatas, por ejemplo, considerando solo las funciones cuadráticas o combinaciones lineales de funciones bases haciendo variar solamente los coeficientes.

Ahora piense en mayor detalle en cómo el concepto de optimización variacional puede ser aplicado al problema de inferencia. Suponga que se tiene un modelo completamente Bayesiano en el cual todos los parámetros están dados por distribuciones a priori. El modelo puede también tener variables latentes así como parámetros, y se debe denotar todo este conjunto por z. Similarmente, se nombra el conjunto de variables observadas por x. El modelo probabilístico representa la distribución conjunta $p(x, z)$, y la meta es encontrar una aproximación para la distribución posterior $p(z|x)$ al igual que para el modelo de evidencia $p(x)$. La deducción mostrada a continuación, es hecha para el caso discreto. Es posible usar el mismo procedimiento para el caso continuo tan solo remplazando las sumatorias por integrales. Para esto, se expresa la verosimilitud del modelo incompleto $p(x)$ como se muestra en la ecuación A.2.

$$\log p(x) = \mathcal{L}(q) + KL(q, p) \tag{A.2}$$

Donde:

$$\mathcal{L}(q) = \sum_z \; q(z) \log \frac{p(x, z)}{q(z)} \tag{A.3}$$

$$KL(q, p) = -\sum_z \; q(z) \log \frac{p(z|x)}{q(z)} \tag{A.4}$$

La ecuación A.2 se cumple debido a que se es posible calcular la verosimilitud del modelo completo mediante la regla del producto como en la ecuación A.5.

$$p(x, z) = p(z|x)p(x) \tag{A.5}$$

Entonces substituyendo A.5 en A.2 se obtiene la ecuación A.6, la cual demuestra que dada cualquier distribución de probalidad $p(z)$, es posible hacer una evaluación de la verosimilitud de modelo incompleto $p(x)$. Por otro lado, esta ecuación también permite plantear el problema mediante una optimización de un funcional que depende de q. Debido a que la distancia $KL(q,p)$ mide la diferencia que existe entre las distribucion $q(z)$ y $p(z|x)$ y a que $KL(q,p) \geq 0$, entonces $KL(q,p) = 0$ si y solo si $q(z) = p(z|x)$. Es decir, el problema de aproximar la distribución $p(z|x)$ puede ser obtenida al minimizar la distancia $KL(q,p)$ entre $q(z)$ y $p(z|x)$ y dicho punto mínimo ocurrira cuando se tenga que $q(z) = p(z|x)$.

$$
\begin{aligned}
\log p(x) &= \mathcal{L}(q) + KL(q,p) \\
&= \sum_z q(z) \log \frac{p(x,z)}{q(z)} - \sum_z q(z) \log \frac{p(z|x)}{q(z)} \\
&= \sum_z q(z) \log \frac{p(z|x)p(x)}{q(z)} - \sum_z q(z) \log \frac{p(z|x)}{q(z)} \\
&= \sum_z q(z) \left(\log \frac{p(z|x)}{q(z)} + \log p(x) \right) - \sum_z q(z) \log \frac{p(z|x)}{q(z)} \\
&= \sum_z q(z) \log \frac{p(z|x)}{q(z)} + \sum_z q(z) \log p(x) - \sum_z q(z) \log \frac{p(z|x)}{q(z)} \\
&= \sum_z q(z) \log p(x) \\
&= \left(\sum_z q(z) \right) \log p(x) \\
&= (1) \log p(x) \\
&= \log p(x)
\end{aligned}
\tag{A.6}
$$

Desde luego que hacer la optimización considerando todas las posibles distribuciones de probabilidad candidatas para $q(z)$ resulta en práctica imposible. Además dicha distribución resulta tan compleja que se vuelve intratable. Por esta razón, se restringe el rango de funciones a trabajar a un conjunto de familias de distribuciones que sea tratables y que aproximen a la verdadera distribución $p(z|x)$, reduciendo así el espacio de búsqueda para la solución. Esta simplificación también resulta útil cuando la familia de distribuciones candidatas tienen formas expecíficas de factorizarse.

Otra manera de restringir la familia de distribuciones que aproximan a $q(z|x)$, es definir para las variables latentes distribuciones a priori gobernadas por un

conjunto de parámetros ω. Así, tanto $\mathcal{L}(q)$ y $KL(q,p)$ dependen directamente de ω, permitiendo el uso de las técnicas estandar de optimización no lineal para encontrar el punto óptimo de ω.

El caso específico en el cual se asume que la distribución de probabilidad factoriza como una partición de variables latentes z en z_1,z_2,...,z_M tales que

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i|\mathbf{x}) \tag{A.7}$$

Donde la mejor distribución q_j^* para cada uno de los factores q_j de la ecuación A.7 está dada por la ecuación A.8.

$$q_j^*(\mathbf{z}_j|\mathbf{x}) = \frac{e^{\mathrm{E}_{i\neq j}[\ln p(\mathbf{z},\mathbf{x})]}}{\int e^{\mathrm{E}_{i\neq j}[\ln p(\mathbf{z},\mathbf{x})]}\, d\mathbf{z}_j} \tag{A.8}$$

y $e^{\mathrm{E}_{i\neq j}[\ln p(\mathbf{z},\mathbf{x})]}$ es la expectación de la distribución conjunta de probabilidad de las variables latentes y los datos, se conoce con el nombre de método variacional Bayesiano.

Apéndice B

Recopilación de resúmenes para clasificación y agrupamiento

B.1. Agrupamiento

Este apéndice contiene partes de los resumenes de los documentos seleccionados para los cálculos en las aplicaciones ejemplificadas en la sección 4.6.4. Para facilitar la visualización, cada palabra ha sido sido etiqueda en el extremo superior derecho con el número perteneciente a su tópico.

B.1.1. Grupo 1

Transforming[47] *growth*[47] *factor*[47]*-beta*[47] *(TGF*[47]*-beta*[47]*) regulates*[47] *many aspects*[32] *of cellular*[12] *function*[47]*. The signaling*[47] *pathway*[47] *is common*[6] *in diverse*[47] *animal*[47] *species*[46] *from vertebrates*[6] *to C*[38]*. elegans*[6]*. Recently*[45] *BMP*[47] *receptor*[47] *associated*[6] *molecule*[38] *2*[45] *(BRAM2*[47]*) was isolated*[6] *from human*[6] *placenta*[49] *cDNA*[6] *library*[17] *by yeast*[6] *two-hybrid*[20] *screening*[6] *and found to bind*[47] *the intracellular*[47] *domain*[6] *of type*[47] *I*[47] *receptor*[47] *(Kurozumi*[2] *et*[2] *al*[17]*., manuscript*[38] *in preparation*[49]*). By data*[6]*- base*[17] *search*[6]*, we found a homologous*[6] *gene*[6] *in C*[47]*. elegans*[17] *which have a 57 %*[6] *amino*[6] *acid*[6] *identity*[6] *over the carboxyl*[6]*-terminal*[49] *60*[6] *amino*[6] *acids*[47] *of BRAM2*[38]*. Tentatively*[6] *we named*[38] *this gene*[47] *as CEBRAM2*[47]*. Full length*[6] *cDNA*[17] *(0*[47]*.9*[17] *Kb*[47]*) of CEBRAM2*[17] *containing*[49] *the SL*[45]*-1*[6] *sequence*[6] *was cloned*[47] *and characterized*[17]*. Northern*[17] *blot*[49] *analysis*[6] *demonstrated*[6] *that the gene*[6] *is expressed*[3] *in all developmental*[38] *stages*[45] *but most strongly*[12] *in embryo*[38]*. GFP*[38] *fusion*[17] *gene*[38] *expression*[47] *under control*[47] *of the CEBRAM2*[38] *promoter*[17] *was analyzed*[38] *to determine*[47] *the cellular*[47] *specificity*[47] *of CEBRAM2*[38] *expression*[19]*. Interestingly*[38] *GFP*[38] *was expressed*[25] *in multiple*[25] *neurons*[38] *in the head*[25] *mostly amphid*[38] *neurons*[12] *(e*[46]*. g*[25]*. ASI*[38]*, ASK*[47]*, and etc.).*

Figura B.1: Parte del resumen perteneciente al documento 1.

The molecular[47] *controls*[27] *governing*[47] *the formation*[4] *of organ*[4] *structure*[4] *during development*[47] *are not well understood*[47]. *The Caenorhabditis*[4] *elegans*[4] *excretory*[4] *cell*[4] *extends*[4] *tubular*[4] *processes*[47], *called*[4] *canals*[4], *along the basolateral*[4] *surface*[4] *of the hypodermis*[4] *and mutations*[24] *in the exc*[4]*-5*[29] *gene*[6] *cause*[47] *tubulocystic*[4] *defects*[29] *at the distal*[32] *tips*[4] *of the canals*[4]. *Here we report*[32] *that exc*[4]*-5*[2] *encodes*[6] *a protein*[6] *homologous*[6] *to guanine*[47] *nucleotide*[47] *exchange*[47] *factors*[47] *(GEFs*[47]*) and contains*[6], *in order*[14], *a Dbl*[47]*/Pleckstrin*[47] *homology*[4] *(DH*[47]*/PH*[47]*) domain*[47], *a cysteine*[6]*-rich*[6] *FYVE*[47] *domain*[6] *and a second*[29] *PH*[47] *element*[27]. *This motif*[27] *architecture*[4] *is similar*[45] *to that of FGD1*[4], *which is responsible*[4] *for faciogenital*[4] *dysplasia*[4] *or Aarskog*[2]*-Scott*[14] *syndrome*[26] *1*[45] . *Ultrastructural*[4] *analysis*[45] *suggests*[4] *that EXC*[4]*-5*[47] *is required*[4] *for the proper*[4] *placement*[4] *of cytoskeletal*[47] *elements*[4] *at the apical*[4] *epithelial*[4] *surface*[4]. *exc*[4]*-5*[47] *interacts*[47] *genetically*[47] *with mig*[47]*-2*[47] *encoding*[47] *a Rho*[47] *GTPase*[47]. *Overexpression*[4] *of exc*[4]*-5*[47] *rescues*[45] *the apical*[4] *defect*[45] *but causes*[4] *defects*[32] *at the basolateral*[4] *surface*[4] *of the excretory*[4] *cell*[47]. *These results*[47] *suggest*[4] *that EXC*[4]*-5*[4] *controls*[47] *the structural*[14] *organization*[32] *of epithelia*[4] *by regulating*[47] *Rho*[47] *family*[27] *GTPase*[47] *activity*[47]. *1*[45] *Pasteris*[4], *N*[6]. *G*[4]. *et*[2] *al*[2]. , *Cell*[47], *79*[2], *669*[2]*-678*[4]*(1994*[2]*)*

Figura B.2: Parte del resumen perteneciente al documento 79.

The bone[47] morphogenetic[47] proteins[47] (BMPs[47]) are a group[47] of transforming[47] growth[47] factor[47] beta[47] (TGF[47]-beta[46])-related[6] factors[47] whose only receptor[47] identified[47] to date[46] is the product[46] of the daf[47]-4[46] gene[38] from Caenorhabditis[46] elegans[6]. Mouse[46] embryonic[47] NIH[47] 3T3[47] fibroblasts[47] display[45] high[46]-affinity[46] 125I[46]-BMP[47]-4[47] binding[47] sites[46]. Binding[46] assays[46] are not possible[6] with the isoform[6] 125I[46]-BMP[47]-2[46] unless[46] the positively[45] charged[6] N[46]-terminal[6] sequence[6] is removed[46] to create[38] a modified[46] BMP[47]-2[45], 125I[46]-DR[46]-BMP[47]-2[2]. Cross[46]-competition[46] experiments[38] reveal[6] that BMP[47]-2[47] and BMP[47]-4[47] interact[47] with the same binding[46] sites[46]. Affinity[46] cross[46]-linking[47] assays[38] show that both BMPs[47] interact[47] with cell[47] surface[47] proteins[6] corresponding[6] in size[47] to the type[46] I (57[6]- to 62[2]-kDa[46]) and type[46] II[47] (75[2]- to 82[46]-kDa[46]) receptor[47] components[47] for TGF[47]-beta[47] and activin[47]. Using[6] a PCR[6] approach[38], we have cloned[6] a cDNA[6] from NIH[47] 3T3[47] cells[47] which encodes[6] a novel[47] member[47] of the transmembrane[6] serine[47]/threonine[47] kinase[47] family[6] most closely[6] resembling[47] the cloned[47] type[47] I receptors[47] for TGF[47]-beta[47] and activin[47].

Figura B.3: Parte del resumen perteneciente al documento 85.

Based[2] on similarities[6] in phenotypes[3] and genetic[9] interactions[6], five Daf[3]-c[3] (dauer[3] constitutive[3]) genes[16], daf[3]-1[47], -4[9], -7[9], -8[2] and -14[19] are thought[6] to have related[47] functions[47]. Molecular[47] identities[6] of four of these genes[16] have been reported[17] previously[2] by Don[3] Riddle[3]'s[2] group[47]. daf[3]-7[9] encodes[9] a homolog[47] of TGF[3]-beta[47](1[2]), daf[3]-1[6] and daf[3]-4[9] encode[6] homologs[16] of TGF[47]-beta[47] receptors[47](2[2], 3[9]), and daf[3]-8[9] encodes[6] a homolog[6] of Drosophila[6] gene[6] Mad[47] (Mothers[47] against dpp[47])(4[2]). We cloned[6] daf[3]-14[19] in order[38] to understand[6] its function[3] in the pathway[47]. daf[3]-14[9] is rescued[9] by the cosmid[6] F01G10[47], recently[6] sequenced[6] by the genome[16] sequencing[6] project[16]. From the sequence[6], we identified[1] a candidate[9] gene[6] based[6] on homology[6], which was confirmed[16] to be daf[3]-14[19] by sequencing[6] mutant[19] alleles[34]. daf[3]-14[19] is a member[47] of the recently[2] described[2] gene[16] family[6](5[9]) that includes[6] Mad[47] from Drosophila[47](6[9]) and the C[6]. elegans[47] genes[16] sma[47]-2[3], sma[47]-3[9], sma[47]-4[9](5[47]) and daf[3]-8[9](4[9]). All of these genes[3] are implicated[6] in TGF[47]-beta[3] related[3] signal[3] transduction[47], suggesting[47] they play[47] a conserved[6] role[19]. Known[19] members[16] of this gene[6] family[6] contain[6] two conserved[47] regions[6], DH1[47] and DH2[47](5[9]). Genefinder[6] analysis[16] and sequence[6] alignment[6] of the daf[3]-14[2] genomic[17] DNA[6] predicts[6] a protein[6] with strong[6] similarity[6] to the DH2[47] region[6] but without a DH1[47]-like[3] domain[47]. Also, no homology[6] to DH1[47] was detected[38] in a search[6] of genomic[17] sequence[6] upstream[47] of daf[3]-14[3] DH2[47] region[6].

Figura B.4: Parte del resumen perteneciente al documento 95.

B.1.2. Grupo 6

The reproductive[32] system is a central[22] regulator[45] of aging[3] in C[22]. elegans[22]. Laser[12] ablation[32] of germline[32] precursors[12] produces[12] a striking[22] extension[3] in lifespan[3]. However, for germline[32] ablation[32] to confer[27] longevity[3], the presence[3] of a normal[32] somatic[12] gonad[32] is essential[32]. These data[27] suggest[27] that the reproductive[3] system produces[32] two kinds[11] of counterbalancing[3] molecular[22] cues[3]: the germline[32] provides[22] signals[3] that diminish[21] lifespan[3] while the somatic[32] gonad[32] produces[12] signal[32](s[11]) that enhance[3] lifespan[3]. We have used[32] a combination[22] of laser[12] ablations[12] and RNAi[16] screening[16] to identify[45] both the cells[12] of the somatic[32] gonad[32] that produce[12] longevity[3] signals[12], and the genes[16] involved[3] in this pathway[45]. Cells[32] of the somatic[32] gonad[32] that promote[27] longevity[3]: The somatic[32] gonad[32] is made up of multiple[48] cell[32] types[27]. To identify[16] the specific[27] cells[32] that produce[32] longevity[3]-promoting[32] signals[32], we devised[44] a 'Twin[32] Ablation[32]' scheme[12] that involves[3] elimination[22] of germ[32] cells[32], followed[12] by elimination[12] of precursors[12] of individual[22] somatic[32] gonad[32] structures[4].

Figura B.5: Parte del resumen perteneciente al documento 6.

AMP[3]-activated[32] protein[8] kinase[37] (AMPK[3]) is activated[3] by high[41] AMP[3]/ATP[41] ratio[7] when the energy[41] level[22] is low[18]. Thus AMPK[3] functions[22] as an energy[41] sensor[41] that couples[3] energy[41] status[22] to metabolism[3], proliferation[32] and survival[41] of the cell[12]. We reason[14] that an organism[22] requires[8] AMPK[3] to survive[18] when AMP[3]/ATP[41] ratio[7] is high[41], for instance[35] during starvation[18]. However, the physiological[22] function[8] and the detailed[18] mechanisms[22] regulating[3] AMPK[3] in an organism[22] during starvation[3] are not fully[18] understood[22]. We found that mutants[3] of aak[3]-2[41], a C[3]. elegans[32] AMPK[3], are sensitive[18] to starvation[3]. Different[12] durations[12] of starvation[3] cause[12] different[8] phenotypes[18] in aak[3]-2[3] mutants[18]; short[3]-term[41] starvation[3] (3[2] days[41] as L1[18]) induces[8] sterility[18] after worms[18] have grown[18] to adulthood[3], and long[3]-term[41] starvation[3] (10[1] days[18] as L1[18]) causes[18] lethality[37] as L1[18]. The sterility[32] is caused[18] by aberrant[37] cell[12] divisions[12] in the germ[32] line[32]. Previous[1] findings[22] show that aak[3]-2[3] mutants[3] continue[18] division[12] of the cell[12] lineages[12] that are normally[18] arrested[18] during L1[3] arrest[18](1[8]).

Figura B.6: Parte del resumen perteneciente al documento 18.

Several genes[3] *that affect*[41] *life*[41] *span*[3] *in C*[10]*. elegans*[3] *act*[3] *in a common*[10] *signalling*[3] *pathway*[3] *that shows*[41] *homology*[10] *to the mammalian*[3] *insulin*[3] *and IGF*[3]*-1*[41] *signalling*[3] *pathways*[3] *and also controls*[41] *dauer*[3] *formation*[3] *in C*[10]*. elegans*[41]*. Both constitutive*[3] *dauer*[3] *formation*[3] *and extended*[41] *life*[41] *span*[41] *of these mutants*[41] *can be suppressed*[3] *by mutation*[3] *in the daf*[3]*-16*[3] *gene*[41]*, which encodes*[3] *a forkhead*[3] *transcription*[3] *factor*[3] *that is inactive*[38] *and resides*[41] *in the cytoplasm*[41] *when phosphorylated*[8] *by the Ins*[3]*/IGF*[3]*-1*[41] *signal*[3]*, and relocates*[41] *to the nucleus*[3] *and controls*[3] *transcription*[3] *of a life*[41] *maintenance*[41] *program*[3] *when dephosphorylated*[10]*. Longevity*[41] *mutants*[3] *with reduced*[41] *activities*[41] *of the Ins*[3]*/IGF*[3]*-1*[41] *pathway*[3] *were recently*[41] *discovered*[10] *in Drosophila*[10] *and mice*[41]*, suggesting*[3] *that this pathway*[3] *for life*[41] *span*[41] *control*[38] *is evolutionary*[41] *conserved*[10]*. A nutritionally*[41] *complete*[10]*, but calorie*[41] *restricted*[41]*, diet*[41] *can significantly*[41] *extend*[41] *life*[41] *span*[41] *in many species*[10]*, pointing*[3] *to another conserved*[10] *mechanism*[48] *of life*[41]*-span*[41] *determination*[41]*.*

Figura B.7: Parte del resumen perteneciente al documento 23.

Dauer[3] *formation*[27] *in C*[3]*. elegans*[38] *is strongly*[3] *induced*[3] *at 27&deg*[3]*;, a temperature*[9] *just*[3] *below the highest*[27] *temperature*[35] *that permits*[3] *growth*[3] *and reproduction*[3]*. Induction*[35] *of dauer*[3] *formation*[3] *at 27&deg*[3]*; occurs*[3] *both in wild*[9]*-type*[35] *strains*[3] *and in sensitized*[35] *mutant*[1] *backgrounds*[1]*. For example*[3]*, unc*[3]*-31*[1] *and unc*[30]*-64*[3] *mutants*[33] *are not Daf*[3]*-c*[3] *at 25&deg*[3]*;, but are strongly*[3] *Daf*[33]*-c*[9] *at 27&deg*[33]*;. unc*[3]*-31*[3] *and unc*[3]*-64*[9] *encode*[30] *proteins*[3] *involved*[33] *in regulated*[3] *secretion*[1] *and the mutants*[33] *have multiple*[3] *behavioral*[3] *phenotypes*[3]*. However, it is not known*[3] *how they affect*[33] *dauer*[33] *formation*[1] *and fit*[3] *into the previously*[1] *characterized*[3] *genetic*[3] *and molecular*[35] *pathways*[3] *regulating*[1] *dauer*[3] *formation*[1]*. We have shown*[3] *that the Daf*[3]*-c*[3] *phenotypes*[3] *of unc*[1]*-31*[3] *and unc*[3]*-64*[3] *mutants*[30] *are suppressed*[3] *by mutations*[30] *in daf*[3]*-16*[3] *but not by mutations*[3] *in daf*[3]*-5*[3]*. These epistasis*[3] *results*[3] *are similar*[3] *to those seen*[9] *for the Daf*[3]*-c*[3] *genes*[1] *daf*[9]*-2*[3] *and age*[33]*-1*[1]*, which define*[3] *an insulin*[3]*-receptor*[3]*/PI3*[1] *kinase*[1] *signaling*[3] *pathway*[3] *that regulates*[3] *both dauer*[3] *formation*[3] *and lifespan*[1]*. Other components*[27] *of this signaling*[3] *pathway*[3] *have not been identified*[3] *genetically*[3] *in C*[3]*. elegans*[3]*.*

Figura B.8: Parte del resumen perteneciente al documento 82.

B.1.3. Grupo 11

pes[12]*-10*[47] *RNA*[31] *and protein*[31] *are expressed*[38] *transiently*[31] *in each somatic*[12] *lineage*[12] *in the pre*[49]*-gastrulation*[12] *embryo*[12] *Geraldine*[12] *Seydoux*[31] *and Andy*[38] *Fire. Carnegie*[12] *Instituhon*[4], *Baltimore*[2] *MD*[2] *21210*[2].

Figura B.9: Resumen perteneciente el documento 11.

Perlecan[4], *a component*[42] *of the extracellular*[4] *matrix*[4] *(ECM*[4]*), is essential*[42] *for myofilament*[42] *formation*[4] *and muscle*[11] *attachment*[42] *in Caenorhabditis*[42] *elegans*[42]. *We show here that perlecan*[42] *is a product*[12] *of muscle*[42] *and that it behaves*[12] *in a cell*[11] *autonomous*[11] *fashion*[11]. *That is, perlecan*[42] *expressed*[4] *in an individual*[11] *muscle*[42] *cell*[12] *does*[11] *not spread*[11] *beyond the borders*[11] *of the ECM*[4] *underlying*[4] *that cell*[12]. *Using*[35] *a polyclonal*[11] *antibody*[42] *that recognizes*[11] *all isoforms*[42] *of perlecan*[42], *we demonstrate*[22] *that this protein*[42] *first appears*[12] *extracellularly*[42] *at the comma*[11] *stage*[12] *(approx*[42]. *350*[12] *min*[12]*) of development*[12]. *We also show that during morphogenesis*[4] *muscle*[42] *cells*[12] *have a heretofore*[22] *undescribed*[10] *plasticity*[22] *of shape*[11]. *This ability*[22] *to regulate*[22] *cell*[12] *shape*[4] *allows*[22] *cells*[12] *within a muscle*[42] *quadrant*[11] *to compensate*[12] *for missing*[11] *cells*[12] *and to form*[11] *a functional*[22] *quadrant*[11]. *A dramatic*[11] *example*[12] *of this morphological*[11] *flexibility*[22] *can be observed*[11] *in animals*[11] *in which the D*[12] *blastomere*[12] *has been removed*[12] *by laser*[11] *ablation*[12]. *Such animals*[42], *lacking*[18] *20*[39] *of the 81*[12] *embryonic*[12] *body*[42] *wall*[42] *muscle*[42] *cells*[22], *can survive*[18] *to become viable*[18] *adult*[4] *animals*[11] *indistinguishable*[46] *from wildtype*[12] *animals*[42].

Figura B.10: Parte del resumen perteneciente al documento 28.

Using[1] *the screen*[1] *for maternal*[12] *effect*[18] *lethal*[18] *mutants*[1] *developed*[12] *by Ken*[18] *Kemphues*[37] *and Jim*[1] *Priess*[12], *we have been looking*[12] *for mutants*[1] *whose inviable*[18] *progeny*[18] *fail*[1] *to make*[1] *gut*[12] *granules*[12]. *With this screen*[1] *we hope*[1] *to identify*[1] *mutations*[1] *in genes*[1] *that are required*[12] *for proper*[37] *specification*[12] *of the E*[12] *lineage*[12]. *The mutants*[1] *that we retain*[11] *from the screen*[1] *have inviable*[18] *progeny*[18] *that divide*[12] *to* > *200*[1] *cells*[12] *(approx*[12].*), are not multinucleate*[11], *and do not make*[12] *gut*[12] *granules*[37] *as detected*[12] *under polarized*[37] *light*[1]. *The mutants*[18] *that we recover*[1] *from this screen*[1] *have fallen*[12] *into three classes*[1]*: par*[37] *mutants*[1] *(described*[1] *by Kemphues*[37] *et*[2] *al*[2].*); those that we are unofficially*[12] *calling*[1] *gut*[12] *mutants*[1] *(gut*[12] *defective*[1]*); and those that we are unofficially*[12] *calling*[1] *&*[12] *'l*[12] *(gut*[12] *granuleless*[11]*). We have concentrated*[1] *on characterizing*[1] *mutants*[12] *of the gut*[11] *and, to some extent*[12], *the ggl*[1] *classes*[1]. *So far*[1] *we have recovered*[18] *16*[12] *gut*[1] *mutants*[12] *and 4*[12] *ggl*[1] *mutants*[12]. *The set*[12] *of gut*[1] *mutants*[1] *contains*[1] *mutations*[2] *in at least 10*[1] *different*[1] *complementation*[1] *groups*[16] *with only three genes*[1] *represented*[1] *by more than one allele*[12]. *The ggl*[1] *class*[1] *of mutant*[1] *is represented*[1] *by four alleles*[1] *in a single*[1] *complementation*[1] *group*[16].

Figura B.11: Parte del resumen perteneciente al documento 34.

We are interested[38] *in understanding*[22] *mesodermal*[12] *patterning*[12] *and fate*[12] *specification*[12] *by studying*[27] *the C*[45]*. elegans*[22] *postembryonic*[39] *mesodermal*[12] *lineage*[12]*, the M*[46] *lineage*[12]*. The M*[2] *lineage*[12] *is derived*[12] *from a single*[45] *precursor*[12] *cell*[12]*, the M*[12] *mesoblast*[12]*, and gives*[12] *rise*[12] *to six cell*[22] *types*[12]*: striated*[42] *bodywall*[42] *muscles*[42] *(BWMs*[42]*), nonmuscle*[42] *coelomocytes*[38] *(CCs*[12]*), and four classes*[22] *of non*[42]*-striated*[42] *sex*[7] *muscles*[42] *which are descendants*[12] *of the sex*[39] *myoblasts*[39] *(SMs*[39]*). We are studying*[27] *the function*[45] *of the mls*[12]*-2*[45] *(mesodermal*[12] *lineage*[12] *specification*[12]*) gene*[27] *in M*[46] *lineage*[12] *patterning*[27] *and fate*[12] *specification*[12]*. The mls*[12]*-2*[45]*(cc615*[12]*) mutation*[45] *causes*[45] *randomization*[12] *of division*[12] *planes*[12] *in the M*[2] *lineage*[12]*, and subsequent*[38] *fate*[12] *transformation*[12] *of CCs*[12] *and BWMs*[42] *to SMs*[39]*. In addition*[38]*, cc615mutants*[12] *have defects*[45] *in SM*[39] *migration*[39] *and show some larval*[39] *and adult*[38] *lethality*[7]*. We have cloned*[38] *the wild*[42] *type*[43] *mls*[12]*-2*[45] *gene*[45] *(C39E6*[12]*.4*[9]*). mls*[12]*-2*[27] *encodes*[45] *a homeodomain*[27] *protein*[20] *that belongs*[27] *to the HMX*[12] *family*[27] *of homeodomain*[27] *proteins*[20] *that are also present*[38] *in sea*[7] *urchin*[12]*, Drosophila*[22] *and vertebrates*[27]*., We examined*[12] *the expression*[27] *pattern*[12] *of mls*[12]*-2*[45] *using*[27] *both functional*[38] *mls*[12]*-2*[45]*::gfp*[38] *fusion*[38] *construct*[38] *and affinity*[46] *purified*[46] *anti*[38]*-MLS*[12]*-2*[45] *antibodies*[38]*.*

Figura B.12: Parte del resumen perteneciente al documento 93.

B.1.4. Grupo 29

A SAGE[16] library[16] was prepared[44] from hand[36]-dissected[44] intestines[21] from adult[4] Caenorhabditis[21] elegans[27], allowing[36] the identification[36] of &gt[21];4000[36] intestinally[27]-expressed[16] genes[27]; this gene[27] inventory[16] provides[36] fundamental[16] information[36] for understanding[16] intestine[27] function[27], structure[36] and development[27]. Intestinally[27]-expressed[27] genes[16] fall[16] into two broad[27] classes[16]: widely[16]-expressed[16] "housekeeping[16]"genes[16] and genes[27] that are either intestine[27]-specific[16] or significantly[16] intestine[16]-enriched[16]. Within this latter class[16] of genes[16], we identified[16] a subset[16] of highly[27]-expressed[16] highly[16]-validated[36] genes[16] that are expressed[27] either exclusively[16] or primarily[34] in the intestine[21]. Over half[36] of the encoded[16] proteins[46] are candidates[16] for secretion[4] into the intestinal[27] lumen[4] to hydrolyze[46] the bacterial[21] food[3] (e[27].g[16]. lysozymes[21], amoebapores[27], lipases[21] and especially[36] proteases[46]). The promoters[16] of this subset[16] of intestine[16]-specific[27]/intestine[27]-enriched[16] genes[16] were analyzed[16] computationally[16], using[16] both a word[36]-counting[44] method[36] (RSAT[12] oligo[16]-analysis[16]) and a method[16] based[36] on Gibbs[27] sampling[16] (MotifSampler[27]). Both methods[36] returned[44] the same over-represented[16] site[27], namely an extended[24] GATA[27]-related[16] sequence[16] of the general[36] form[12] AHTGATAARR[21],

Figura B.13: Parte del resumen perteneciente al documento 29.

Gene[27] *regulatory*[27] *networks*[22] *that control*[27] *the terminally*[12] *differentiated*[27] *state*[31] *of a cell*[12] *are, by and large*[16], *only superficially*[25] *understood*[31]. *In a mutant*[45] *screen*[27] *aimed*[22] *at identifying*[1] *regulators*[27] *of gene*[27] *batteries*[27] *that define*[22] *the differentiated*[12] *state*[31] *of two left*[12]*/right*[25] *asymmetric*[12] *C*[6]. *elegans*[15] *gustatory*[25] *neurons*[25], *ASEL*[25] *and ASER*[25], *we have isolated*[1] *a mutant*[25], *fozi*[42]*-1*[45], *with a novel*[42] *mixed*[46]*-fate*[27] *phenotype*[25], *characterized*[6] *by de-repression*[31] *of ASEL*[25] *fate*[12] *in ASER*[25]. *fozi*[42]*-1*[45] *codes*[6] *for a protein*[6] *that functions*[27] *in the nucleus*[12] *of ASER*[25] *to inhibit*[31] *the expression*[27] *of the LIM*[25] *homeobox*[27] *gene*[27] *lim*[27]*-6*[25], *neuropeptide*[46]*-encoding*[6] *genes*[16] *and putative*[16] *chemoreceptors*[25] *of the GCY*[25] *gene*[27] *family*[6]. *The FOZI*[42]*-1*[45] *protein*[6] *displays*[46] *a highly*[6] *unusual*[6] *domain*[27] *architecture*[27], *that combines*[6] *two functionally*[27] *essential*[6] *C2H2*[27] *zinc*[27]*-finger*[27] *domains*[6], *which are probably*[6] *involved*[1] *in transcriptional*[27] *regulation*[27], *with a formin*[42] *homology*[16] *2*[25] *(FH2*[42]*) domain*[6], *normally*[25] *found only in cytosolic*[46] *regulators*[31] *of the actin*[42] *cytoskeleton*[42]. *We demonstrate*[25] *that the FH2*[42] *domain*[6] *of FOZI*[42]*-1*[45] *has lost*[25] *its actin*[42] *polymerization*[42] *function*[45] *but maintains*[27] *its phylogenetically*[6] *ancient*[22] *ability*[27] *to homodimerize*[27]. *fozi*[42]*-1*[45] *genetically*[22] *interacts*[9] *with several transcription*[27] *factors*[27] *and micro*[22] *RNAs*[31] *in the context*[22] *of specific*[31] *regulatory*[27] *network*[22] *motifs*[27].

Figura B.14: Parte del resumen perteneciente al documento 60.

Multi[36]*-cellular*[16] *organisms*[36] *needs*[36] *appropriate*[36] *regulatory*[27] *systems*[36] *which control*[27] *to activate*[27]*/in-activate*[27] *transcriptional*[27] *process*[36] *of multiple*[16] *genes*[16] *at proper*[27] *stages*[16] *and in proper*[27] *cells*[27] *for development*[27]*. In many cases*[36]*, this is regulated*[16] *through the binding*[27] *of proteins*[27] *to a specific*[27] *region*[27] *of the gene*[16]*. Such protein*[27] *binding*[27] *sites*[27] *are known*[16] *as cis*[27]*-regulatory*[27] *elements*[27] *or motifs*[27]*. However, despite*[27] *their hypothetical*[49] *importance*[27]*, cell*[27]*- and stage*[18]*-specific*[27] *regulatory*[27] *motifs*[27] *in multi*[36]*-cellular*[16] *organisms*[36] *remain*[49] *largely*[16] *unrevealed*[16]*. Moreover, the prediction*[36] *of motifs*[16] *by in silico*[16] *methods*[36] *and the verification*[36] *of putative*[49] *motifs*[27] *by experimental*[35] *methods*[36] *are both challenging*[36] *problems*[36]*. To address*[36] *these problems*[36]*, we developed*[36] *a new*[16] *computer algorithm*[36] *for extracting*[36] *cell*[13]*-/stage*[16]*-specific*[16] *regulatory*[27] *motifs*[27] *of C*[87]*. elegans*[16] *genes*[16]*. Since the size*[35] *and position*[49] *of such motif*[27] *are not known*[27] *before analysis*[36]*, we have to search*[36] *short*[49] *sequences*[49] *(5*[2] *bases*[49] *or so) in long*[49] *target*[27] *sequence*[16] *area*[36] *(more(than 1000*[36] *bases*[49]*), which cause*[34] *too many pseudo*[36] *positives*[16] *and too long*[36] *computation*[36] *time*[36]*. Thus, we developed*[36] *an algorithm*[36] *named*[49] *&quotfiltering*[36] *step&quot*[16] *which reduces*[36] *the search*[36] *space*[16] *(and(therefore pseudo*[36] *positives*[36]*) dramatically*[16]*, without losing*[49] *the real*[36] *positives*[16]*.*

Figura B.15: Parte del resumen perteneciente al documento 66.

GENBANK[6] *-M93256*[7] *The tra*[7] *-1*[31] *gene*[27] *of Caenorhabditis*[6] *elegans*[20] *is a major*[27] *developmental*[27] *regulator*[31] *that promotes*[31] *female*[7] *development*[27]*. Two mRNAs*[31] *are expressed*[6] *from the tra*[7] *-1*[31] *locus*[27] *as a result*[34] *of alternative*[6] *mRNA*[31] *processing*[31]*. One mRNA*[31] *encodes*[6] *a protein*[6] *with five zinc*[27] *fingers*[27] *and the other a protein*[6] *with only the first two zinc*[31] *fingers*[31]*. We have derived*[12] *a preferred*[27] *in vitro*[6] *DNA*[27] *binding*[6] *site*[6] *for the five finger*[6] *protein*[31] *by selection*[27] *from random*[35] *oligonucleotides*[27]*. The two finger*[27] *protein*[6] *does*[27] *not bind*[6] *to DNA*[27] *in vitro*[6]*. Moreover, removal*[31] *of the first two fingers*[31] *from the five finger*[31] *protein*[31] *does*[27] *not eliminate*[7] *binding*[6] *and has little*[31] *effect*[31] *on its preferred*[27] *binding*[27] *site*[34]*. We find that a protein*[20] *sequence*[6] *amino*[6] *-terminal*[6] *to the finger*[27] *domain*[6] *also appears*[27] *to play*[6] *a role*[14] *in DNA*[27] *binding*[27]*.*

Figura B.16: Resumen perteneciente al documento 91.

B.2. Recuperación

En este apéndice se presentan los resúmenes de los documentos 19,20,55,34 y 93 que representan a los primeros 5 documentos más relevantes para la consulta realizada con las palabras "asymmetric" y "cell" en la colección de 100 documentos tomados como muestra de la base de datos original.

Las ocurrencias de las palabras han sido marcadas para una fácil identificación, señalando la palabra "asymmetric" con un superíndice [1] y la palabra "cell" con un superíndice [2].

Por cuestiones de espacio, en ocasiones no es posible proporcionar el texto completo de los documentos. Por lo tanto, para solventar este problema se han tenido que suprimir ciertos segmentos del resumen considerados como no relevantes. Cuando esto ocurre, el contenido de dicho segmento sera reemplazado por "..." indicando que en esa posición cierta parte del resumen ha sido suprimida.

Es importante recordar al lector que la probabilidad de aparición para cualquier palabra en un tópico y para la aparición de un tópico en un documento, nunca es cero. Esto ocurre debido a lo explicado en la sección 4.2, donde el parámetro de suavizado otorga a todas las combinaciones una probabilidad muy pequeña. Por este motivo debe quedar claro que el que una palabra no aparezca en un documento, no significa necesariamente que esta palabra tenga probilidad cero de aparicón en los tópicos del documento, sino que simplemente esta no fue muestrada.

Un ejemplo de este comportamiento se aprecia en la figura B.19, donde no se visualiza la palabra "asymmetric", pero la probabilidad $\sum_{j=1}^{K} p(w_k|z_{i,j} = 1)p(z_{i,j} = 1|d_i)$ no es cero en ningún caso, por lo que tampoco ocurre que la ecuación 4.31 sea cero en algún momento.

Por otro lado, debe recordarse que una frecuencia de aparición alta, si representa una probabilidad de aparición elevada, ya que los estimadores para los parámetros son calculados de forma proporcional a dicha frecuencia.

Por este motivo se puede observar que en el documento 19 de la figura B.17 las palabras de consulta se presentan en el texto con altas probabilidades de aparición.

Asymmetric[1] cell[2] division is a fundamental process that produces celler diversity during development. In C. elegans, asymmetric divisions of certain blast cells, including the T blast cell[2], are regurated by lin-17/frizzled and lin-44/wnt. It has been proposed that the LIN-44 signal, which acts through the LIN-17 receptor, provides polarity to cells that undergo asymmetric division. To make clear this model, we expressed lin-44 ectopically, and examined effects on asymmetric cell[2] division. In normal development, the anterior daughter of the T cell[2] produces hypodermal cells, and the posterior daughter produces neural cells.

...

As a result of expression of lin-44 at the anterior of the T cell[2] in lin-44 mutants, polarity reversal phenotype was greatly enhanced (97Moreover the anterior expression reverses the polarity of the division even in wild type (14These results demostrate that direction of cell[2] polarity is controled by the LIN-44 signal. Although pop-1/tcf has been shown to be required for asymmetric T cell[2] division, involvement of -catenin has not been shown. We found that wrm-1/-catenin mutants were defective in the asymmetric T cell[2] division as observed in lin-17 mutants. This suggests that the asymmetric cell[2] division is controled by -catenin in the canonical Wnt pathway.

Figura B.17: Parte del resumen perteneciente al documento 19.

Asymmetric[1] cell[2] division depends on coordinating the position of the mitotic spindle with the axis of cellular polarity. In C. elegans embryos, the initial cellular polarity is established through the asymmetric localization of PAR proteins, which subsequently regulates the asymmetric distribution of the cell[2] fate determinants and spindle positioning. However, it is still unclear how spindle positioning is coordinated with the PAR polarity cues. We provide evidence that LET-99 is a link between PAR polarity cues and the downstream machinery that determines spindle positioning in C. elegans embryos. In let-99 1-cell[2] embryos the nuclear-centrosome complex exhibits a hyperactive oscillation that is dynein-dependent, instead of the normal anteriorly-directed migration and rotation of the nuclear-centrosome complex. Further, at anaphase in let-99 embryos the spindle poles do not show the characteristic asymmetric movements typical of wild type. LET-99 is a DEP domain protein that is asymmetrically enriched in a band that encircles P lineage cells. The LET-99 localization pattern is dependent on PAR-3 and PAR-2 and correlates with nuclear rotation and anaphase spindle pole movements in wild-type embryos, as well as with changes in these movements in par mutant embryos. In particular, LET-99 is uniformly localized in 1-cell[2] par-3 embryos at the time of nuclear rotation. Rotation fails in spherical par-3 embryos where the eggshell has been removed, but rotation occurs normally in spherical wild-type embryos.

Figura B.18: Resumen perteneciente al documento 26.

mig-13 is a guidance factor that promotes cell[2] migrations in the anterior direction (Sym et. al., 1999). Previous work demonstrated that mig-13 is required for the anterior migrations of the QR descendants and the BDU neurons (Sym et. al., 1999). Consistent with the role of mig-13 in anterior migrations, we have also found that mig-13 also directs the anterior migration of the distal tip cell[2] (DTC) in the posterior gonad arm during late L3. We are taking several approaches to understand how mig-13 can guide many anterior migrations. mig-13 encodes a novel transmembrane protein containing putative protein-protein interaction domains: a CUB domain and a LDL-receptor repeat in the extracellular region as well as a proline-rich domain in the intracellular region (Sym et. al., 1999). We have examined the function of these domains in MIG-13 by deleting them and assaying the in vivo activity of the resulting MIG-13 construct. Our data suggests that a MIG-13 construct lacking the intracellular domain can confer partial function in directing the QR descendants to the anterior. Previous mosaic analysis revealed that mig-13 acts non-autonomously to direct the migrations of the QR lineage (Sym et. al., 1999). To determine where mig-13 expression is sufficient to guide the migrating cells, we have expressed mig-13 in different sets of tissues, as well as in specific subsets of cells.

Figura B.19: Parte del resumen perteneciente al documento 55.

Using the screen for maternal effect lethal mutants developed by Ken Kemphues and Jim Priess, we have been looking for mutants whose inviable progeny fail to make gut granules.

...

Embryos of gut mutants divide in a normal asymmetric[2] *and asynchronous pattern during very early cell*[2] *divisions, thus distinguishing them from par mutants. P granules are also segregated normally. Embryonic cell*[2] *divisions in gut embryos appear to be normal until the beginning of gastrulation. At that time in wild-type development the two E cells' division rate slows and the E cells begin to move into the center of the embryo. In gut mutants the E cells' division rate does not slow and the E cells do not gastrulate. These embryos arrest development embryonically as a ball of cells (approx. 400-600) with no apparent morphogenesis. As well as lacking gut granules, the embryos fail to produce two antigens that are normally found in differentiated gut cells (assayed by antibodies J126 and SP37 from S. Strome). They do undergo some differentiation; we observe cell*[2]*- death nuclei, neuronal nuclei, large amounts of MyoB (anti-myoB from D. Miller) and hypodermal antigens (gift of MH27 from R. Waterston). In embryos of ggl mutants early divisions including E cell*[2] *division and gastrulation appear to be normal.*

...

Figura B.20: Parte del resumen perteneciente al documento 34.

We are interested in understanding mesodermal patterning and fate specification by studying the C. elegans postembryonic mesodermal lineage, the M lineage. The M lineage is derived from a single precursor cell[2]*, the M mesoblast, and gives rise to six cell*[2] *types: striated bodywall muscles (BWMs), nonmuscle coelomocytes (CCs), and four classes of non-striated sex muscles which are descendants of the sex myoblasts (SMs). We are studying the function of the mls-2 (mesodermal lineage specification) gene in M lineage patterning and fate specification. The mls-2(cc615) mutation causes randomization of division planes in the M lineage, and subsequent fate transformation of CCs and BWMs to SMs. In addition, cc615mutants have defects in SM migration and show some larval and adult lethality. We have cloned the wild type mls-2 gene (C39E6.4). mls-2 encodes a homeodomain protein that belongs to the HMX family of homeodomain proteins that are also present in sea urchin, Drosophila and vertebrates.*

...

Figura B.21: Parte del resumen perteneciente al documento 93.

Apéndice C

Recopilación de datos para el Modelo de Autores y Tópicos

C.1. Títulos por autor para NIPS

Este apéndice contiene una recopilación de los títulos de los primero tres autores más probables para cada tópico en la colección, que aparecen en la tabla 5.3.

Debido a la gran producción académica de algunos autores, es imposible poner todos los documentos contenidos en esta colección; por esta razón, algunos autores presentan al final de la tabla una signo de "...", cuyo significado es denotar que existen más documentos en la colección que por cuestiones de espacio no seran presentados.

C.1.1. Tópico 6

Eric Mjolsness
Neural Networks for Model Matching and Perceptual Organization
A Lagrangian Approach to Fixed Points.
Visual Grammars and their Neural Nets
Clustering with a Domain-Specific Distance Measure
Two-Dimensional Object Localization by Coarse-to-Fine Correlation Matching
LEARNING WITH PREKNOWLEDGE: CLUSTERING WITH FOINT AND GRAPH MATCHING DISTANCE MEASURES
NEW ALGORITHMS FOR 2D AND 3D POINT MATCHING: POSE ESTIMATION AND CORRESPONDENCE
A Multiscale Attentionai Framework for Relaxation Neural Networks
A Convergence Proof for the Softassign Quadratic Assignment Algorithm,
From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data

Anand Rangarajan
Clustering with a Domain-Specific Distance Measure
LEARNING WITH PREKNOWLEDGE: CLUSTERING WITH FOINT AND GRAPH MATCHING DISTANCE MEASURES
NEW ALGORITHMS FOR 2D AND 3D POINT MATCHING: POSE ESTIMATION AND CORRESPONDENCE
Softassign versus Softmax: Benchmarks in Combinatorial Optimization
A Framework for Nonrigid Matching and Correspondence
A Convergence Proof for the Softassign Quadratic Assignment Algorithm

John Platt
Constrained Differential Optimization
Analog Circuits for Conslxained Optimization
Learning by Combining Memorization and Gradient Descent .
Networks for the Separation of Sources that are Superimposed and Delayed
An Analog VLSI Chip for Radial Basis Functions
Postal Address Block Location Using a Convolutional Locator Network
A CONVOLUTIONAL NEURAL NETWORK HAND TRACKER
A Neural Network Classifier for the I1000 OCR Chip
A Constructive RBF Network for Writer Adaptation,
Using Analytic QP and Sparseness to Speed Training of Support Vector Machines,
Large Margin DAGs for Multiclass Classification,
Support Vector Method for Novelty Detection

C.1.2. Tópico 16

Michael I. Jordan
REINFORCEMENT LEARNING ALGORITHM FOR PARTIALLY OBSERVABLE MARKOV DECISION PROBLEMS
REINFORCEMENT LEARNING W1TH SOFT STATE AGGREGATION
BOLTZMANN CHAINS AND HIDDEN MARKOV MODELS
AN ALTERNATIVE MODEL FOR MIXTURES OF EXPERTS
ACTIVE LEARNING wrrH STATISTICAL MODELS
COMPUTATIONAL STRUCTURE OF COORDINATE TRANSFORMATIONS: A GENERALIZATION STUDY
Factorial Hidden Markov Models
Exploiting Tractable Substructures in Intractable Networks
Fast Learning by Bounding Likelihoods in Sigmoid Type Belief Networks
Learning Fine Motion by Markov Mixtures of Experts
Reinforcement Learning by Probability Matching
A Variational Principle for Model-based Morphing,
Recursive Algorithms for Approximating Probabilities in Graphical Models,
Hidden Markov Decision Trees
...

Zoubin Ghahramani
Supervised Learning from Incomplete Data via an EM Approach
FORWARD DYNAMIC MODELS IN HUMAN MOTOR CONTROL: PSYCHOPHYSICAL EVIDENCE
FACTORIAL LEARNING AND THE EM ALGORITHM
ACTIVE LEARNING wrrH STATISTICAL MODELS
COMPUTATIONAL STRUCTURE OF COORDINATE TRANSFORMATIONS: A GENERALIZATION STUDY
Factorial Hidden Markov Models
Hidden Markov Decision Trees,
Hierarchical Non-linear Factor Analysis and Topographic Maps,
Learning Nonlinear Dynamical Systems Using an EM Algorithm,
SMEM Algorithm for Mixture Models,
Variational Inference for Bayesian Mixtures of Factor AnaIysers,
Learning to Parse Images,

Volker Tresp
A Neural Network Approach for Three-Dimensional Object Recognition
Neural Control for Rolling Mills: Incorporating Domain Theories to Overcome Data Deficiency
Some Solutions to the Missing Feature Problem in Vision
Network Structuring and Training Using Rule-based Knowledge
Training Neural Networks with Deficient Data
COMBINING ESTIMATORS USING NON-CONSTANT WEIGHTING FUNCTIONS
EFFICIENT METHODS FOR DEALING WITH MISSING DATA IN SUPERVISED LEARNING
Discovering Structure in Continuous Variables Using Bayesian Networks
Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Terms and Network Averaging
Early Brain Damage,
Nonlinear Markov Networks for Continuous Variables,
A Solution for Missing Data in Recurrent Neural Networks with an Application to Blood Glucose Prediction,
...

Dean Pomerleau
ALVINN: An Autonomous Land Vehicle in a Neural Network
Rapidly Adapting Artificial Neural Networks for Autonomous Navigation
Input Reconstruction Reliability Estimation
Non-Intrusive Gaze Tracking Using Artificial Neural Networks
USING A SALIENCY MAP FOR ACTIVE SPATIAL SUBJECTIVE ATTENTION: IMPLEMENTATION & INITIAL RESULTS
A CONNECIIONIST TECHNIQUE FOR ACCELERATED TEXTUAL INPUT: LETTING A NETWORK DO THE TYPING

Shumeet Baluja
Non-Intrusive Gaze Tracking Using Artificial Neural Networks
USING A SALIENCY MAP FOR ACTIVE SPATIAL SUBJECTIVE ATTENTION: IMPLEMENTATION &INITIAL RESULTS
Human Face Detection in Visual Scenes
Using the Future to "Sort Out"the Present: Rankprop and Multitask Learning for Medical Risk Evaluation
Genetic Algorithms and Explicit Search Statistics,
Using Expectation to Guide Processing: A Study of Three Real-World Applications,
Making Templates Rotationally Invariant: An Application to Rotated Digit Recognition,
Probabilistic Modeling for Face Orientation Discrimination: Learning from Labeled and Unlabeled Data,

Clay Spence
Neuronal Maps for Sensory-Motor Control in the Barn Owl
The Computation of Sound Source Elevation in the Barn Owl
Applications of Neural Networks in Video Signal Processing
COARSE-TO-FINE IMAGE SEARCH USING NEURAL NETWORKS
Applications of Multi-Resolution Neural Networks to Mammography,
Hierarchical Image Probability (HIP) Models,
Unmixing Hyperspectral Data,

C.1.3. Tópico 31

Richard P. Lippmann
Neural Net and Traditional Classifiers
Adaptive Neural Net Preprocessing for Signal Detection in Non-Gaussian Noise
Practical Characteristics of Neural Network and Conventional Pattern Classifiers on Artificial and Speech Problems
HMM Speech Recognition with Neural Net Discrimination
Using Genetic Algorithms to Improve Pattern Classification Performance.
Practical Characteristics of Neural Network and Conventional Pattern Classifiers.
Improved Hidden Markov Model Speech Recognition Using Radial Basis Function Networks
A Boundary Hunting Radial Basis Function Classifier which Allocates Centers Constructively
Figure of Merit Training for Detection and Spotting
USING VOICE TRANSFORMATIONS TO CREATE ADDITIONAL TRAINING TALKERS FOR WORD SPOTTING
PREDICTING THE RISK OF COMPLICATIONS IN CORONARY ARTERY BYPASS OPERATIONS USING NEURAL NETWORKS
A Micropower Analog VLSI HMM State Decoder for Wordspotting,

Eric I. Chang
Using Genetic Algorithms to Improve Pattern Classification Performance .
A Boundary Hunting Radial Basis Function Classifier which Allocates Centers Constructively
Figure of Merit Training for Detection and Spotting
USING VOICE TRANSFORMATIONS TO CREATE ADDITIONAL TRAINING TALKERS FOR WORD SPOTTING

Dietrich Wettschereck
Improving the Performance of Radial Basis Function Networks by Learning Center Locations
Locally Adaptive Nearest Neighbor Algorithms

C.1.4. Tópico 44

Chris Williams
Directional-Unit Boltzmann Machines
USING A NEURAL NETTO INSTANTlATE A DEFORMABLE MODEL
EM Optimization of Latent-Variable Density Models
Computing with Infinite Networks,
Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo,
GTM: A Principled Alternative to the Self- Organizing Map,
Regression with Input-dependent Noise: A Gaussian Process Treatment,
Finite-Dimensional Approximation of Gaussian Processes,
Discovering Hidden Features with Gaussian Processes Regression,
DTs: Dynamic Trees,
Adding Constrained Discontinuities to Gaussian Process Models of Wind Fields,
A MCMC Approach to Hierarchical Mixture Modelling,
...

Christopher M. Bishop
ESTIMATING CONDITIONAL PROBABILITY DENSroES FOR PERIODIC VARIABLES
REAL-TIME CONTROL OF TOKAMAK PLASMA USING NEURAL NETWORKS
EM Optimization of Latent-Variable Density Models
Bayesian Model Comparison by Monte Carlo Chaining,
Regression with Input-Dependent Noise: A Bayesian Treatment,
GTM: A Principled Alternative to the Self- Organizing Map,
Ensemble Learning for Multi-Layer Networks,
Approximating Posterior Distributions in Belief Networks Using Mixtures,
Regression with Input-dependent Noise: A Gaussian Process Treatment,
Bayesian PCA,

David Barber
Online Learning from Finite Training Sets: An Analytical Case Study,
Bayesian Model Comparison by Monte Carlo Chaining,
Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo,
On-line Learning from Finite Training Sets in Nonlinear Networks,
Ensemble Learning for Multi-Layer Networks,
Radial Basis Functions: A Bayesian Treatment,
Tractable Variational Structures for Approximating Graphical Models,
Gaussian Fields for Approximate Inference in Layered Sigmoid Belief Networks,

C.1.5. Tópico 50

Naftali Tishby
Information, Prediction, and Query by Committee
Statistical Modeling of Cell-Assemblies Activities in Associative Cortex of Behaving Monkeys
The Power of Amnesia
The Statistical Mechanics of k-Satisfaction
Decoding Cursive Scripts
Agnostic Classification of Markovian Sequences,
Synergy and Redundancy among Brain Cells of Behaving Monkeys,
Multi-Electrode Spike Sorting by Clustering Transfer Functions,
Information Capacity and Robustness of Stochastic Neuron Models,
AggIomerative Information Bottleneck,

Michael Kearns
Estimating Average-Case Learning Curves Using Bayesian, Statistical Physics and VC Dimension Methods
A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-test Split
Inference in Multilayer Networks via Large Deviation Bounds,
Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms,
Reinforcement Learning for Spoken Dialogue Systems,
Approximate Planning in Large POMDPs via Reusable Trajectories,

Sumio Watanabe
An Optimization Method of Layered Neural Networks Based on the Modified Information Criterion
Solvable Models of Artificial Neural Networks
Algebraic Analysis for Non-regular Learning Machines,

C.2. Títulos por autor para wormbase

C.2.1. Tópico 3

Lemire BD
Mitochondrial respiratory chain deficiency in Caenorhabditis elegans results in developmental arrest and increased life span.
Stable heteroplasmy but differential inheritance of large mitochondrial DNA deletion in nematodes.
Mitochondrial ATP synthase controls larval developmental cell nonautonomously in Caenorhabditis elegans.
The role of mitochondria in the life of the nematode
The mitochondrial prohibitin complex is essential for embryonic viability and germline function in Caenorhabditis elegans.
The ubiquinone-binding site of the Saccharomyces cerevisiae succinate-ubiquinone oxidoreductase is a source of superoxide.
Mitochondrial complex I mutations in Caenorhabditis elegans produce cytochrome c oxidase deficiency, oxidative stress and vitamin-responsive lactic acidosis.
Caenorhabditis elegans development requires mitochondrial function in the nervous system.
...

Clarcke CF
A dietary source of coenzyme Q is essential for growth of long-lived Caenorhabditis elegans clk-1 mutants.
Extension of life-span in Caenorhabditis elegans by a diet lacking coenzyme Q.
Development and fertility in Caenorhabditis elegans clk-1 mutants depend upon transport of dietary coenzyme Q8 to mitochondria.
Reproductive fitness and quinone content of Caenorhabditis elegans clk-1 mutants fed coenzyme Q isoforms of varying length.
Yeast and rat Coq3 and Escherichia coli UbiG polypeptides catalyze both O-methyltransferase steps in coenzyme Q biosynthesis.
Ubiquinone biosynthesis in Saccharomyces cerevisiae. Isolation and sequence of COQ3, the 3,4-dihydroxy-5-hexaprenylbenzoate methyltransferase gene.
A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the Escherichia coli ubiE gene.
Coenzyme Q and aging in the nematode Caenorhabditis elegans.
respiratory deficiency in Saccharomyces cerevisiae abc1 mutants.
Yeast COQ4 encodes a mitochondrial protein required for coenzyme Q synthesis.
...

Watts JL
par-6, a gene involved in the establishment of asymmetry in early C. elegans embryos, mediates the asymmetric localization of PAR-3.
Isolation and characterization of a delta(5)-fatty acid desaturase from Caenorhabditis elegans.
The C. elegans par-4 gene encodes a putative serine-threonine kinase required for establishing embryonic asymmetry.
Identification and characterization of an animal Delta(12) fatty acid desaturase gene by heterologous expression in Saccharomyces cerevisiae.
A palmitoyl-CoA-specific Delta 9 fatty acid desaturase from Caenorhabditis elegans.
Genetic dissection of polyunsaturated fatty acid synthesis in Caenorhabditis elegans.
Polyunsaturated fatty acid synthesis: what will they think of next?
zu170 Defines a New Gene, par-6, and Can Act as a Suppressor of par-2
The par-4 gene encodes a protein kinase that is cortically distributed in early embryos
...

C.2.2. Tópico 18

Hall HD
Genetics of cell and axon migrations in C. elegans.
The unc-5, unc-6, and unc-40 genes guide circumferential migrations of pioneer axons and mesodermal cells on the epidermis in C. elegans.
The posterior nervous system of the nematode Caenorhabditis elegans: Serial reconstruction of identified neurons and complete pattern of synaptic interactions.
Kinesin-related gene unc-104 is required for axonal transport of synaptic vesicles in C. elegans.
Freeze-fracture and freeze-etch studies of the nematode Caenorhabditis elegans.
Motor vesicles.
Combinatorial control of touch receptor neuron expression in Caenorhabditis elegans.
Mutations in the Caenorhabditis elegans beta-tubulin gene mec-7: Effects on microtubule assembly and stability and on tubulin autoregulation.
Electron microscopy and three-dimensional image reconstruction.
The mab-21 gene of Caenorhabditis elegans encodes a novel protein required for choice of alternate cell fates.
Homologies in the neurogenesis of nematodes, arthropods and chordates.
...

Hardin JD
MEX-3 is a KH domain protein that regulates blastomere identity in early C. elegans embryos.
An actin-mediated two-step mechanism is required for ventral enclosure of the C. elegans hypodermis.
The VAB-1 Eph receptor tyrosine kinase functions in neural and epithelial morphogenesis in C. elegans.
A putative catenin-cadherin system mediates morphogenesis of the Caenorhabditis elegans embryo.
Cytokinesis and midzone microtubule organization in Caenorhabditis elegans require the kinesin-like protein ZEN-4.
Dynamics and ultrastructure of developmental cell fusions in the Caenorhabditis elegans hypodermis.
The cellular mechanism of epithelial rearrangement during morphogenesis of the Caenorhabditis elegans dorsal hypodermis.
Rapid epithelial-sheet sealing in the Caenorhabditis elegans embryo requires cadherin-dependent filopodial priming.
A degrading way to make an organ.
Cell lineage analysis. Videomicroscopy techniques.
Getting into shape: epidermal morphogenesis in Caenorhabditis elegans embyros.
...

Hedgecock EM
The mating system of C. elegans: Evolutionary equilibrium between self- and cross-fertilization in a facultative hermaphrodite.
Normal and mutant thermotaxis in the nematode C. elegans.
A gene required for nuclear and mitochondrial attachment in the nematode C. elegans.
Mutations affecting programmed cell deaths in the nematode C. elegans.
Polyploid tissues in the nematode C. elegans.
Cell lineage mutants in the nematode C. elegans.
Axonal guidance mutants of Caenorhabditis elegans identified by filling sensory neurons with fluorescein dyes.
Mutant sensory cilia in the nematode C. elegans.
Transposon tagging of genes affecting axonal outgrowth in C. elegans.
Genetics of cell and axon migrations in C. elegans.
The unc-5, unc-6, and unc-40 genes guide circumferential migrations of pioneer axons and mesodermal cells on the epidermis in C. elegans.
Motor vesicles.
...

C.2.3. Tópico 29

Horvitz HR
Nondisjunction mutants of the nematode C. elegans.
A uniform genetic nomenclature for the nematode C. elegans.
Post-embryonic cell lineages of the nematode, C. elegans.
Laser microbeam techniques in biological research.
Genetic analysis of C. elegans.
unc-93(e1500): A behavioral mutant of C. elegans that defines a gene with a wild-type null phenotype.
Isolation and genetic characterization of cell-lineage mutants of the nematode C. elegans.
Abnormal cell lineages in mutants of the nematode C. elegans.
Gonadal cell lineages of the nematode Panagrellus redivivus and implications for evolution by the modification of cell lineage.
Nematode postembryonic cell lineages.
Serotonin and octopamine in the nematode C. elegans.
Postembryonic nongonadal cell lineages of the nematode Panagrellus redivivus: Description and comparison with those of C. elegans.
Programmed cell death in nematode development.
Factors that influence neural development in nematodes.
Neurone differentiation in cell lineage mutants of C. elegans.
...

Watts JL
TWO MODIFIER SCREENS FOR NEW GENES INVOLVED IN PROGRAMMED CELL DEATH
FUNCTIONAL STUDIES OF THE CLASS B SYNMUVS, GENES REQUIRED FOR NEGATIVE REGULATION OF VULVAL INDUCTION, AND CHARACTERIZATION OF THE CLASS B SYNMUV GENE lin-61
A SCREEN FOR ESSENTIAL CELL-DEATH GENES
A mod-5 SUPPRESSION SCREEN FOR GENES INVOLVED IN SEROTONERGIC NEUROTRANSMISSION
A SCREEN FOR GENES SYNTHETICALLY LETHAL WITH lin-35 Rb
A NOVEL PHENOTYPE OF TRANSGENE MISEXPRESSION YIELDS NEW INSIGHT INTO THE SYNMUV GENES
A NEW CLASS OF SYNMUV MUTATIONS IS PREDICTED TO DISRUPT A HISTONE ACETYLTRANSFERASE COMPLEX
SQV-4 UDP-GLUCOSE DEHYDROGENASE IS TEMPORALLY AND SPATIALLY REGULATED TO CONTROL C. ELEGANS VULVAL MORPHOGENESIS
A GENETIC PATHWAY FOR THE CONTROL OF THE SEXUALLY DIMORPHIC DEATHS OF THE CEM NEURONS
...

Watts JL
Genetic analysis of defecation in Caenorhabditis elegans.
Cell interactions coordinate the development of the C. elegans egg-laying system.
Chemosensory cell function in the behavior and development of Caenorhabditis elegans.
Genetic analysis of chemosensory control of dauer formation in Caenorhabditis elegans.
Cell interactions involved in development of the bilaterally symmetrical intestinal valve cells during embryogenesis in Caenorhabditis elegans.
Evidence for parallel processing of sensory information controlling dauer formation in Caenorhabditis elegans.
Chemosensory regulation of development in C. elegans.
Thinking about genetic redundancy.
Sequence of C. elegans lag-2 reveals a cell-signalling domain shared with Delta and Serrate of Drosophila.
A screen for nonconditional dauer-constitutive mutations in Caenorhabditis elegans.
Regulation of a periodic motor program in C. elegans.
The mind of a worm.
...

C.2.4. Tópico 32

Edgley ML
Genetic Analysis in Caenorhabditis elegans.
Genetic balancers.
Two pleiotropic classes of daf-2 mutation affect larval arrest, adult behavior, reproduction and longevity in Caenorhabditis elegans.
LG II balancer chromosomes in Caenorhabditis elegans: mT1(II III) and the mIn1 set of dominantly and recessively marked inversions.
Systematic interactome mapping and genetic perturbation analysis of a C. elegans TGF-beta signaling network.
mC6: The Full Monty
the CGC catalog.
New Versions of the CGC Bibliography, Strain List and Map Data
On the Impending Transfer of Genetics Center Services
Call for Worms
A Genetic Toolkit for C. elegans
Deletions Targeted on unc-101
New Balancers for LG II
Wild-Type Chromosomes Carrying GFP Markers
freezing and thawing worms.
...

Durbin RM
A survey of expressed genes in Caenorhabditis elegans.
2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans.
The genome of the nematode Caenorhabditis elegans.
A workbench for large-scale sequence homology analysis.
ACeDB and Macace.
The C. elegans expression pattern database - A beginning.
Transposon Tc1-derived, sequence-tagged sites in Caenorhabditis elegans as markers for gene mapping.
Pfam - A comprehensive database of protein domain families based on seed alignments.
Tc7, a Tc1-hitch hiking transposon in Caenorhabditis elegans.
Analysis of protein domain families in Caenorhabditis elegans.
Gene expression and development databases for C. elegans.
WormBase: network access to the genome and biology of Caenorhabditis elegans.
Systematic gene inactivation in C. elegans
RNA sequence analysis using covariance models.
A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.
Genome Project Database
Genome Sequencing
...

Schatz BR
Automatic thesaurus generation for an electronic community system.
The Worm Community System, Release 2.0 (WCSr2).
Pattern discovery in gene regulation designing an analysis environment.
A concept space approach to addressing the vocabulary problem in scientific information retrieval - An experiment on the Worm Community System.
The Worm Community System
The Electronic Gazette
Availability of WCS (Worm Community System), Release 1
Worm Community System Users Envision Future Application.
THE WORM COMMUNITY SYSTEM.
WCS and Mosaic: The ENQUire System

C.2.5. Tópico 48

Ruvkun GB
The C. elegans cell lineage and differentiation gene unc-86 encodes a protein with a homeodomain and extended similarity to transcription factors.
The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2 and Caenorhabditis elegans unc-86 gene products.
Molecular genetics of the Caenorhabditis elegans heterochronic gene lin-14.
The Caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch.
Caenorhabditis elegans has scores of homeobox-containing genes.
The unc-86 gene product couples cell lineage and cell identity in C. elegans.
Nematode homeobox cluster.
Dominant gain-of-function mutations that lead to misregulation of the C. elegans heterochronic gene lin-14, and the evolutionary implications of dominant mu
Negative regulatory sequences in the lin-14 3' untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development.
...

Zarkower D
Molecular analysis of the C. elegans sex-determining gene tra-1– a gene encoding 2 zinc finger proteins.
Zinc fingers in sex determination - Only one of the two C. elegans tra-1 proteins binds DNA in vitro.
Regulatory rearrangements and smg-sensitive alleles of the C. elegans sex-determining gene tra-1.
Dominant feminizing mutations implicate protein-protein interactions as the main mode of regulation of the nematode sex-determining gene tra-1.
Evidence for evolutionary conservation of sex-determining genes.
Similarity of DNA binding and transcriptional regulation by Caenorhabditis elegans MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms.
mab-3 is a direct tra-1 target gene regulating diverse aspects of C. elegans male sexual development and behavior.
Direct protein-protein interaction between the intracellular domain of TRA-2 and the transcription factor TRA-1A modulates feminizing activity in C. elegans.
Establishing sexual dimorphism: Conservation amidst diversity?
Polycomb group regulation of Hox gene expression in C. elegans.
Evolutionary conservation of sex determining genes
...

Kaech SM
LET-23 receptor localization by the cell junction protein LIN-7 during C. elegans vulval induction.
The LIN-2LIN-7LIN-10 complex mediates basolateral membrane localization of the C. elegans EGF receptor LET-23 in vulval epithelial cells.
lag-2 is Not Required for the Secondary Cell Fate in Vulval Induction
Protein Interactions Between the Receptor Tyrosine Kinase LET-23, LIN-7 and LIN-2
PROTEIN INTERACTIONS BETWEEN LIN-2, LIN-7 AND THE RECEPTOR TYROSINE KINASE LET-23: A POSSIBLE MECHANISM FOR LOCALIZING LET-23 IN THE PN.P CELLS DURING VULVAL INDUCTION
BASOLATERAL LOCALIZATION OF LET-23 RTK BY A LIN-2LIN-7LIN-10 PROTEIN COMPLEX IN VULVAL PRECURSOR CELLS
CELLULAR JUNCTIONS AND LATERAL SIGNALING IN VULVAL DEVELOPMENT
LIN-2, LIN-7 and LET-23 Protein Interactions Provide a Mechanism for LET-23 Localization during Vulval Development

C.2.6. Tópico 50

Shuichi Onami
antitative cell division pattern analysis on RNAi embryos
Computer Simulations and Objective in vivo Measurements reveal Length-dependent Pulling Force as the Primary Mechanism for Male Pronuclear Migration
Cell Lineage Analysis using Automatic Cell Lineage Extraction System
Automatic cell lineage acquisition system and analysis of early embryogenesis
Cell lineage acquiring system for early embryogenesis
Analysis of early embryonic cell lineage of RNAi-treated embryos using automatic cell lineage acquisition system
Analysis on Migration of the Sperm Pronucleus Assisted by Computer Simulations.
Automatic measurement system for early embryonic cell lineage
Computer-assisted Analyses reveal Length-dependent Pulling Force as the Primary Mechanism for Male Pronuclear Migration
Computer-assisted Analyses of Microtubule-dependent Forces for Centrosome Positioning in One-Cell Embryo.
Functional analysis of poly(A) elongation in translational control of maternal glp-1 mRNA
...

Fire A
Integrative transformation of C. elegans.
Proper expression of myosin genes in transgenic nematodes.
Vectors for low copy transformation of C. elegans.
A modular set of lacZ fusion vectors for studying gene expression in Caenorhabditis elegans.
CeMyoD accumulation defines the body wall muscle cell fate during C. elegans embryogenesis.
Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle.
Body-wall muslce formation in Caenorhabditis elegans embryos that lack the MyoD homolog hlh-1.
The novel metallothionein genes of Caenorhabditis elegans - structural organization and inducible, cell-specific expression.
Molecular characterization of the her-1 gene suggests a direct role in cell signaling during Caenorhabditis elegans sex determination.
Functional conservation of nematode and vertebrate myogenic regulatory factors.
Histochemical techniques for locating Escherichia coli beta-galactosidase activity in transgenic organisms.
Sequence requirements for myosin gene-expression and regulation in Caenorhabditis elegans.
...

White JG
Cell cycling and DNA replication in a mutant blocked in cell division in the nematode C. elegans.
Electron microscopical reconstruction of the anterior sensory anatomy of the nematode C. elegans.
Connectivity changes in a class of motoneurone during the development of a nematode.
The structure of the ventral nerve cord of C. elegans.
Laser microbeam techniques in biological research.
Regulation and cell autonomy during postembryonic development of C. elegans.
On the control of germ cell development in Caenorhabditis elegans.
Neurone differentiation in cell lineage mutants of C. elegans.
The embryonic cell lineage of the nematode C. elegans.
Factors that determine connectivity in the nervous system of C. elegans.
Polyploid tissues in the nematode C. elegans.
Neuronal connectivity in C. elegans.
The structure of the nervous system of the nematode C. elegans.
Determination of cell division axes in the early embryogenesis of C. elegans.
...

Apéndice D

Descripción del Software

En general, la implementación del Software fue realizada a través del Software Matlab. Para el muestreo de Gibbs, se ha utilizado el toolbox llamado "Matlab Topic Modeling Toolbox 1.3.2", el cual puede ser descargado desde `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.

Sin embargo, para las aplicaciones mostradas en este documento se han realizado un nuevo toolbox, que toma la información procesada por las rutinas de muestreo y realiza varias tareas. Entre estas tareas están aquellas donde se procesa el modelo final obtenido y se calculan probabilidades, obtienen distancias y demás insumos necesarios para llegar a un resumen muy concreto de los resultados. Para este tipo de procesamientos matemáticos se han utilizado las herramientas matemáticas que Matlab proporciona en su conjunto de toolboxes estándar.

Por otro, lado el preprocesamiento de las colecciones es también una tarea a resolver y que ha generado software. Debido a que no existe un formato estándar para la manipulación de los textos, fue necesario desarrollar ciertas aplicaciones que tomarán el contenido y lo convierta al formato usado por Matlab Topic Modeling Toolbox 1.3.2. Estas rutinas de software fueron programas en lenguaje C++ y shell script mediante utilerías estándar de unix, tales como grep y tr. El principal requierimiento de este software es contar únicamente con un sistema unix o linux que tenga un compilador ansi de C++. Finalmente, cabe aclarar que el preprocesamiento de las bases de datos es una tarea muy demandante en tiempo de cómputo. Debido a esto, el lenguaje C++ constituye una buena opción ya que ha comprobado ser eficiente en cuestiones de tiempo de ejecución.

Bibliografía

Bishop, C.M. (2007). *Pattern Recognition and Machine Learning*. Springer, 1st edn. 5, 8, 19, 22

Blei, D.M., Ng, A.Y. y Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine learning research JOURNAL OF MACHINE LEARNING RESEARCH*, **3**, 993–1022. 33, 35

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. y Harshman, R.A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**, 391–407. 34

Elango, P.K. y Jayaraman, K. (2005). Clustering images using the latent dirichlet allocation model. 33

Endres, F., Plagemann, C., Stachniss, C. y Burgard, W. (2009). Unsupervised discovery of object classes from range data using latent dirichlet allocation. 33

Erricson, K.A. y Kintsch, W. (1995). Long-term working memory. 49

Feynman, R., Leighton, R. y Sands, M. (1964). *The Feynman Lectures on Physics*, vol. 2. Addison-Wesley, Boston, 2nd edn. 132

Gelman, A. y Rubin, D.B. (1992). Iterative and non-iterative simulation algorithms. *Computing Science and Statistics*, 433–438. 25, 31

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. 30

Geyer, C.J. (1992). Practical markov chain monte carlo (with discussion). *Statis. Sci.*, **7**, 473–511. 25

Gilks, W. y Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman and Hall/CRC, 1st edn. 19

GRIFFITHS, T., STEYVERS, M. Y TENENBAUM, J. (2007). Topics in semantic representation. *Psychological Review*, **114**, 211–244. 46

GRIFFITHS, T.L. Y STEYVERS, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 5228–5235. 42, 44, 50, 55

HEINRICH, G. (2008). Parameter estimation for text analysis,. Tech. rep., University of Leipzig. 42, 44

HOLMES, D. Y FORSYTH, R. (1995). The federalist revisited: New directions in authorship attribition. *Literary and Linguistic Computing*, **10**, 111–127. 86

JORDAN, M.I. (1999). *Learning in graphical models*. The IMT Press. 4

KASS, R.E. Y RAFERTI, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795. 59

KINTSCH, W. (1988). The role of knowledge in discourse comprehesion: A construction-integration model. 49

LAURITZEN, S.L. (1996). *Graphical Models*. Oxford University Press. 11

MANNING, C.D. Y SCHUETZE, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 1st edn. 58

METROPOLIS, N. Y ULAM, S. (1949). The monte carlo method. *Journal of Statictical Association*, **44**, 335–341. 19

MULLER, P. Y QUINTANA, F. (2004). Nonparametric bayesian data analysis. *Statistical Science*, **19**, 95–110. 60

NEWMAN, D., SMYTH, P. Y STEYVERS, M. (2006). Scalable parallel topic models. 125

NEWMAN, D., ASUNCION, A., SMYTH, P. Y WELLING, M. (2008). Distributed inference for latent dirichlet allocation. 125

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann,. 11

POTTER, M.C. (1993). Very short term conceptual memory. 49

ROSEN-ZVI, M., GRIFFITHS, T., STEYVERS, M. Y SMYTH, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487–494, AUAI Press, Arlington, VA, USA. 86

Rosen-zvi, M., Griffiths, T., Steyvers, M. y Smyth, P. (2005). Learning author topic models from text corpora. 86

Steyvers, M. y Griffiths, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates. 51

Steyvers, M., Smyth, P., Zvi, M.R. y Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 306–315, ACM, New York, NY, USA. 86

Thomas, H. (1999). Probabilistic latent semantic indexing. 34

Witten, I.H. y Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann series in data management systems, Morgan Kaufmann, 2nd edn. 2

Yang, Y. (1997). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, **1**, 67–88. 33

Glosario

Expectation Maximization (EM) es un método para encontrar estimaciones de máxima verosimilitud de los parámetros en modelos estadísticos donde se depende de variables latentes o no observadas. EM es un método iterativo que alterna entre la ejecución de un paso de expectativa (E), que calcula la expectativa de la log-verosimilitud con una estimación actual para las variables latentes, y la maximización (paso M), que estima los parámetros que maximizan la espectativa obtenida en el paso E. *página 24*

Explaining Away es un patrón de razonamiento, en el cual la confirmación de una causa de un evento observado reduce la necesidad de recurrir a causas alternativas. Lo opuesto al Explaining Away también ocurre, cuando la confirmación de una causa incrementa la prosibilidad en otra. *página 14*

Análisis de Semántica Latente (LSA) es una técnica usada en el procesamiento de lenguaje natural, en particular en semántica vectorial. Se busca realizar un análisis de la relación entre un conjunto de documentos y los términos que contienen, al producir un conjunto de conceptos relativos a la muestra obtenida. *página 33*

Plate Notation es una forma alternativa para representar variables que se repiten en un modelo de gráficas. En lugar de mostrar cada variable que se repite de forma individual, una placa o rectángulo es usado para agrupar las variables que aparecen de forma reiterativa en una subgráfica. El número de ciclos de aparición es denotado a través de una valor que va colocado dentro de la placa. A partir de ese momento, se hace la asunción de que dicha subgráfica es duplicada tantas veces como el valor dentro de la placa. Las variables dentro de la subgráfica son indexadas por el número, y las aristas que atraviezan a la placa son replicadas por cada repetición del subgrafo. *página 8*

Análisis Probabilístico de la Semántica Latente (PLSA) es una versión probabilística del LSA, que ahora considera un modelo de gráficas probabilísti-

cas. Tiene sus aplicaciones en los campos de recuperación de información y filtrado, procesamiento del lenguaje natural, aprendizade por computadora, etc. *página 34*

Stop Words es el nombre dado a un conjunto de palabras las cuales carecen de significado semántico y por lo tanto no son útiles en tareas de procesamiento del lenguaje natural. *página 64*