UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELAMIENTO Y SEGUIMIENTO DE TÓPICOS PARA DETECCIÓN DE MODUS OPERANDI EN ROBO DE VEHÍCULOS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DIEGO GARRIDO

PROFESOR GUÍA:
RICHARD WEBER

SANTIAGO, CHILE
2021

Dedicado a mis padres: Gabriela y Pedro que sin ellos no habría llegado a ser lo que soy.

Agradecimientos

En primer lugar, quiero dar agradecimiento a mi madre Gabriela, quién me ha acompañado, cuidado y amado incondicionalmente. A mi padre Pedro, por su cariño, preocupación y guía. Gracias por ayudarme a ser una persona de bien, humilde y formar una visión esperanzadora de la vida.

Al profesor Richard, por la oportunidad de iniciarme en el mundo de los datos que tanto me apasiona, por la confianza, apoyo y preocupación.

Gracias!

Tabla de Contenidos

Índice de Tablas

Índice de Ilustraciones

RESUMEN DE LA TESIS PARA OPTAR
AL TÍTULO DE MAGÍSTER EN GESTIÓN
DE OPERACIONES
E INGENIERÍA CIVIL INDUSTRIAL
POR: **DIEGO GARRIDO**
FECHA: 2021
PROF. GUÍA: RICHARD WEBER

MODELAMIENTO Y SEGUIMIENTO DE TÓPICOS PARA DETECCIÓN DE MODUS OPERANDI EN ROBO DE VEHÍCULOS

En este trabajo se describe una metodología para el descubrimiento de tópicos en el tiempo. La metodología propuesta está basada en (i) discretización del corpus en épocas, (ii) descubrimiento de tópicos en cada época mediante Hierarchical Dirichlet Process (HDP), (iii) la construcción de un grafo de similitud entre tópicos de épocas adyacentes, el cual permite modelar cambios entre los tópicos como: nacimiento, muerte, evolución, división y fusión. En contraste a trabajos anteriores, la metodología propuesta utiliza Word Mover's Distance (WMD) como medida de similitud entre tópicos, medida que destaca por ser robusta a tópicos que no poseen un vocabulario común, debido a que trabaja con sus *word embeddings*. Se reportan resultados experimentales tanto cuantitativos como cualitativos en el fenómeno de robo de vehículos en Chile, usando como corpus los relatos de víctimas de robo de vehículo entre los años 2011-2016 provistos por la Asociación de Aseguradores de Chile (AACH). El algoritmo propuesto logra capturar bien los tópicos latentes del corpus, descubriendo delitos tales como robo sin presencia del conductor, robo con violencia y "portonazo".

Capítulo 1

Introducción

Grandes volumenes de datos digitales son alamacenados día a día, en forma de noticias, blogs, páginas web, artículos científicos, libros, imágenes, sonido, video, redes sociales, etc. Volviéndose clave contar con herramientas computacionales que ayuden a organizar, buscar y entender grandes colecciones de datos.

Si pudieramos buscar y explorar documentos en base a sus temas, podríamos enfocar nuestra búsqueda en temas específicos o más amplios, podríamos observar como estos temas cambian en el tiempo o como se relacionan unos a otros. En vez de buscar documentos únicamente a través de palabras claves, podríamos primero hallar temas que son de nuestro interés, y luego examinar los documentos relacionados a ese tema. Por ejemplo, podríamos descubrir nuevas tendencias de investigación, analizar la evolución de la contigencia social, estudiar la efectividad de campañas publicitarias en base a la opinión de los consumidores, organizar y recomendar contenido en un blog, etc.

El objetivo del trabajo de tesis es desarrollar una metodología que permita descubrir tópicos en el tiempo, siendo capaz de modelar cambios tales como: nacimiento, muerte, evolución, división y fusión. Adicionalmente, debe ser robusta a cambios en el vocabulario en el tiempo, permitiendo comparar tópicos de épocas adyacentes a pesar que de no tener un vocabulario común.

1.1. Metodología Propuesta

Los modelos de tópicos probabilísticos ayudan a descubrir los temas latentes (*clusters*) en una colección de documentos, como estos temas están conectados unos a otros y cómo cambian en el tiempo. Permiten resumir un gran colección de documentos a través de sus temas y organizarlos entorno a estos. Estos tratan a un tópico como una distribución de probabilidad discreta sobre el vocabulario de un corpus. Una práctica habitual es interpretar un tópico a partir de sus N palabras más probables. Por ejemplo, para $N = 5$ las palabras más probables de un tópico son: “llaves”, “domicilio”, “individuos”, “casa” y “porton”, por lo que una etiqueta valida para este tópico podría ser “portonazo”.

En esta tesis se propone una metodología para el descubrimiento de tópicos en el tiempo. Esta metodología consiste en la discretización del corpus en épocas, el descubrimiento de

tópicos en cada época mediante Hierarchical Dirichlet Process (HDP), la construcción de un grafo de similitud entre tópicos de épocas adyacentes, el cual permite modelar cambios entre los tópicos como: nacimiento, muerte, evolución, división y fusión. En contraste a trabajos anteriores, la metodología propuesta utiliza Word Mover's Distance (WMD) como medida de similitud entre tópicos. Esta medida destaca por su robustez ante tópicos que no poseen un vocabulario común, debido a que trabaja sobre el espacio de los *word embeddings*.

1.2. Caso de estudio

Se escoge el problema del robo de vehículos o accesorios de vehículos como caso de estudio, debido a que es un problema que afecta a toda la sociedad en Chile y en el mundo, problema que se ha vuelto más relevante el último tiempo debido al crecimiento en el robo de vehículo motorizado y accesorios (ver Figura 1.1).

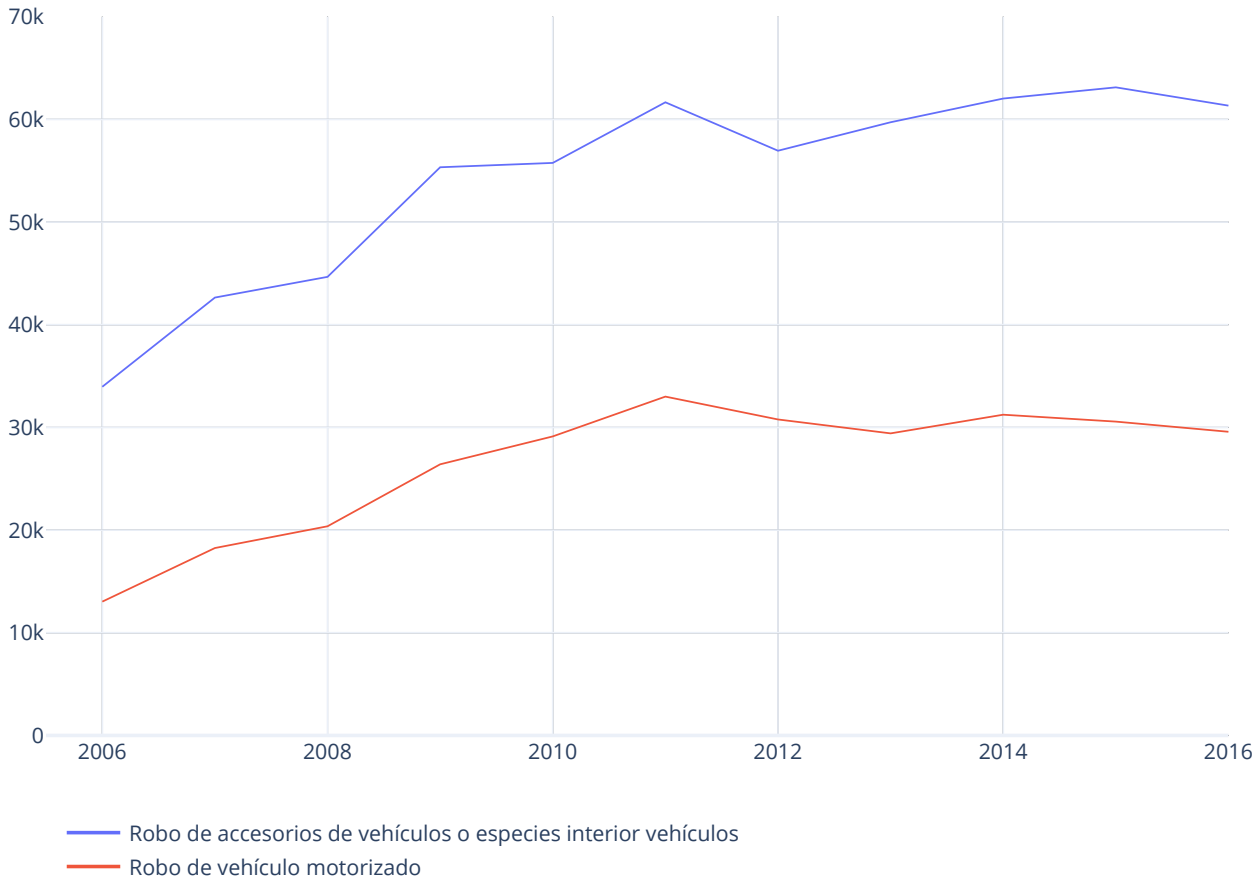


Figura 1.1: Cantidad de robos de vehículos y accesorios anuales en Chile entre los años 2006-2016. Fuente: Informe anual de Carabineros, 2006-2016, INE.

Este fenómeno trae consigo un montón de costos para la sociedad, como incremento en la percepción de la seguridad, aumentos en la prima de los seguros de los asegurados, aumento

en los costos de las aseguradoras [1] y el incremento de otros tipos de delitos. [2]

El corpus utilizado correspode a una colección de 49,015 relatos de víctimas de robo de vehículo, entre los años 2011-2016, provistos por la Asociación de Aseguradores de Chile (AACH). Cabe destacar que se estima que un tercio del parque automotriz se encuentra asegurado[3], por lo que se trabaja con una muestra del parque automotriz.

En el contexto de robo de vehículos, los tópicos vendrían siendo los "modus operandi" que utilizan los delicuentes para robar un vehículo. Así, la metodología propuesta permitiría descubrir los *modus operandi* ocultos en los relatos de las víctimas y caracterizarlos a partir de las palabras, como también ver su evolución a través del tiempo, siendo capaz de detectar cuando nacen y mueren, y como cambian en el tiempo.

1.3. Revisión del estado del arte

El problema enunciado consiste en una tarea de *clustering*, debido a que no se cuenta con una etiqueta del tema al que corresponde cada documento, siendo el propósito del trabajo descubrirla. El modelamiento de tópicos es uno de los enfoques más prometores de clustering aplicado a texto, siendo su objetivo descubrir los temas (*clusters*) ocultos presentes en el corpus, permitiendo resumir, organizar y explorar grandes colecciones de datos.

Algunas de las técnicas de modelamiento de tópicos están basadas en factorización matricial como LSI (Latent Semantic Indexing) (Dumais, 2004) o NMF (Non-negative Matrix Factorization)(Xu et al., 2003), pero este trabajo está basado en modelos probabilísticos generativos, como LDA (Latent Dirichlet Allocation)(Blei et al., 2003) o HDP (Hierarchical Dirichlet Process)(Teh et al., 2005). Ambos enfoques tienen sus pros y contras, en este trabajo se prefiere el enfoque probabilístico ya que es capaz de expresar incertidumbre en la asignación de un tópico a un documento y en la asignación de palabras a los tópicos, además, este enfoque suele aprender tópicos más descriptivos (Stevens et al., 2012).

En el modelamiento de tópicos se pueden presentar los siguientes dinamismos:

1. **Evolución de los tópicos**: la evolución de los tópicos se refleja en el cambio en la distribución sobre las palabras. Por ejemplo, el "portonazo" en un determinado momento se comete en grupos de 2-3 personas con arma blanca, luego evoluciona de arma blanca a arma de fuego y lo perpetran jóvenes menores de edad.

2. **Dinámismo en la mezcla de tópicos**: esto permite capturar la popularidad de los tópicos en el tiempo.

[1] Considerando que el costo promedio incurrido en un auto asegurado robado y no recuperado es de $ 5,000,000 de pesos, la pérdida total considerando solo los vehículos no recuperados para el año 2015 es de unos $15,720 millones de pesos.

[2] El destino de los vehículos robados es variado, se usan los autos para perpetrar otros delitos y huir, venderlos por piezas en talleres clandestinos o blanquear sus documentos para pasarlos por la frontera y venderlos o cambiarlos por droga en el extranjero.

[3] http://www.economiaynegocios.cl/noticias/noticias.asp?id=185224

3. **Nacimiento, muerte, fusión y división de tópicos**: En el contexto de robos es natural que en el tiempo aparezcan nuevos *modus operandi* como también que desaparezcan aquellos que ya no parecen tan atractivos.

En el modelamiento de tópicos estático destaca LDA y HDP. La diferencia principal es que LDA necesita de antemano fijar el número de tópicos a descubrir y HDP lo infiere a partir del corpus.

Dentro de los primeros modelos de tópicos dinámicos exitosos está Dynamic Topic Modelling (DTM)(Blei and Lafferty, 2006) junto Topic Over Time (TOC)(Wang and McCallum, 2006). Estos modelos mantienen el número de tópicos fijo en el tiempo, por lo que si aparece un nuevo tópico este quedará clasificado dentro de un tópico preexistente desde el comienzo, por lo que solo es capaz de capturar el punto 1 y 2.

En (Ahmed and Xing, 2012) se propone Dynamic Hierarchical Dirichlet Process (DHDP), modelo que no mantiene el número de tópicos fijo en el tiempo, sino que lo infiere a partir del corpus. Sin embargo, este modelo no es capaz de capturar fusión y división de tópicos. Además, a diferencia de los otros modelos de tópicos mencionados, DHDP no es una tecnología ampliamente usada y no cuenta con una implementación disponible, por lo que se desconoce su desempeño en otras fuentes de información.

En (Wilson and Robinson, 2011) y (Beykikhoshk et al., 2018) se propone una metodología que permite capturar los dinámismos mencionados utilizando LDA y HDP respectivamente. Estas consisten en dividir el corpus en épocas, entrenar de forma independiente un modelo de tópico en cada época, para finalmente unir los resultados obtenidos. En este trabajo se utilizan técnicas de modelado dinámico de tópicos bajo este enfoque, usando HDP para el descubrimiento de tópicos en cada época.

1.4. Estructura de la tesis

En el capítulo 2 se describen los fundamentos teóricos en los que se basa la metodología propuesta la cual es descrita en el capítulo 3. Luego, en el capítulo 4 se presenta un análisis cuantitativo y cualitativo de la metodología propuesta en el fenómeno de robo de vehículos. Finalmente, en el capítulo 5 se presentan las conclusiones y futuras lineas de investigación.

Capítulo 2

Marco teórico

En este capítulo se describen los conceptos fundamentales para comprender la metodología propuesta en el capitulo 3. El capítulo es estructurado como sigue. En la sección 2.1 se introduce a nivel general los modelos de *clustering* probabilísticos conocidos como *mixture models*. En la sección 2.2 se describen los modelos de tópicos estáticos LDA y HDP. Finalmente, en la sección 2.3 se explica una metodología que modela la evolución de los tópicos en el timepo.

2.1. Mixture Models

Uno de los supuestos básicos en *clustering* es asumir que cada observación x_i pertenece a un solo *cluster* k. Podemos expresar la asignación a un *cluster* como una variable aleatoria z_i, donde $z_i = k$ significa que x_i pertenece al *cluster* k. La variable z_i no es observada en los datos y se considera una variable oculta. Cada *cluster* posee un parámetro ϕ_k que codifica su información. Podemos obtener la distribución que caracteriza a un solo *cluster* k condicionando en z_i

$$p(x_i|z_i = k, \phi) = p(x_i|\phi_k) \tag{2.1}$$
$$\tag{2.2}$$

Además, podemos definir la probabilidad de que una nueva observación pertenezca al *cluster* k

$$p(z_i = k|\pi) = \pi_k \tag{2.3}$$

Con $\sum_k \pi_k = 1$, ya que π_k son probabilidades de eventos mutuamente excluyentes. La distribución de x_i es entonces de la forma

$$p(x_i) = \sum_k p(z_i = k|\pi)p(x_i|z_i = k, \phi) = \sum_k \pi_k p(x_i|\phi_k) \tag{2.4}$$

Podemos escribir $p(x_i|\phi_k)$ como $x_i \sim F(\phi_{z_i})$, donde F es la distribución asociada a las

observaciones.

Se obtiene una representación equivalente del modelo al considerar el parámetro ϕ_{z_i} usado para generar la observación x_i proviene de una distribucion discreta G, la cual tiene la forma

$$G(\phi) = \sum_k \pi_k \delta_{\phi_k}(\phi) \tag{2.5}$$

En otras palabras, G es una mezcla de funciones delta, donde la probabilidad de que ϕ sea igual a ϕ_k es π_k. Luego, un *mixture model* se puede representar como a continuación

$$\phi_{z_i} \sim G \tag{2.6}$$
$$x_i \sim F(\phi_{z_i}) \tag{2.7}$$

Un **Bayesian mixture model** es un *mixture model* con una medida aleatoria para las mezclas. En la sección 2.1.1. y 2.1.2 nos referimos a dos *priors* ampliamente usados para construir *bayesian mixture model*: la distribución Dirichlet que nos permite construir un **finite mixture model**, donde el número de átomos o *clusters* a descubrir es finito, denotado por K y un *prior* no paramétrico denominado Dirichlet Process (DP), el cual permite construir un **infinite mixture model**, donde el número de *clusters* no está acotado.

2.1.1. Distribución Dirichlet

La distribución Dirichlet (Minka, 2000) es una generalización multivariada de la distribución beta, la cual tiene soporte sobre un **simplex**, definido por:

$$S_K = \left\{ x : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1 \right\} \tag{2.8}$$

Luego, su función de densidad de probabilidad (pdf):

$$Dir(x|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(x \in S_K) \tag{2.9}$$

, donde $B(\alpha) = B(\alpha_1, \ldots, \alpha_K)$ es la generalización de la función beta a K variables:

$$B(\alpha) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \tag{2.10}$$

, donde $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$.

En la Figura 2.1 se observa el efecto de los parámetros en la distribución Dirichlet con $K = 3$. El parámetro α_k controla la *sparsity*, mientras más se acerca a 0 los vectores generados tienen más átomos nulos y se concentra la masa en unas pocas coordenadas, mientras más grande α_k la masa más se concentra en el centro (1/3, 1/3, 1/3). Cuando $\alpha_k = 1$ se tiene una distribución uniforme en el dominio S_K. Por otro lado, cuando α no es simétrico la masa

se concentra proporcionalmente en las coordenadas con α_k mayor.

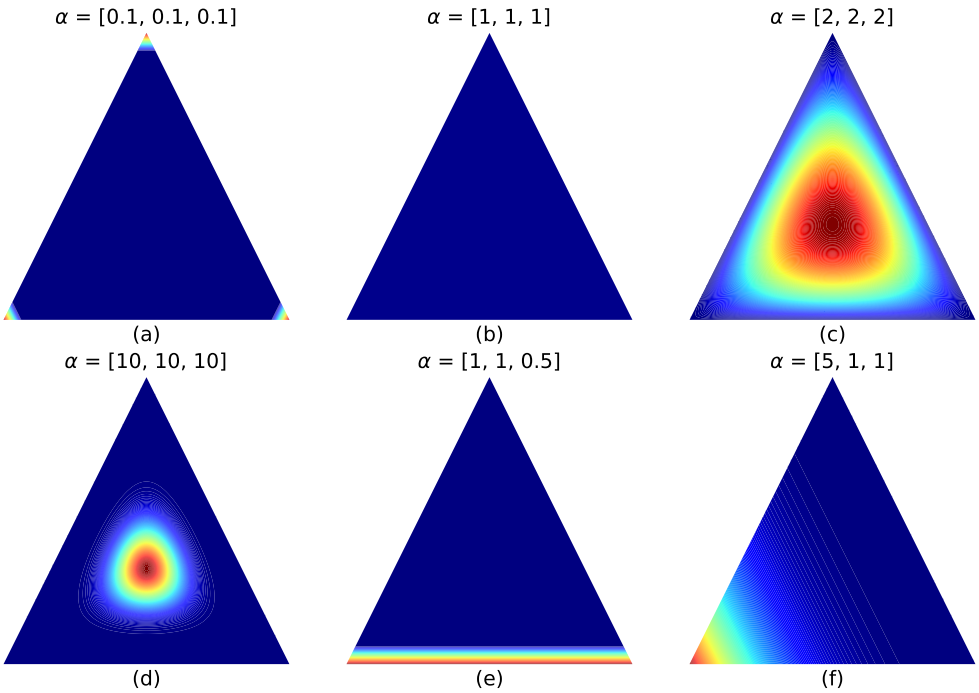


Figura 2.1: Densidad de la distribución Dirichlet con $K = 3$. Define una distribución sobre el *simplex*, el cual puede ser representado por una superficie trinagular.

En general se asume simetría en los parámetros de la distribución Dirichlet de la forma $\alpha_k = \frac{\alpha}{K}$, de esta manera α funciona como parámetro de concentración. En la Figura 2.2 se oberva una realización de una distribución Dirichlet con $\alpha \in \{0.1, 1, 10\}$ y $K \in \{2, 10, 100\}$. En esta figura podemos observar que a mayor α los compenentes del vector x más similares se vuelven, esto es más notorio a mayor dimencionalidad debido a que existen más dimensiones a las que distribuir la masa.

La distribución Dirichlet es comúnmente usada en estadística Bayesiana, debido a que que es *prior* conjugado de la distribución categórica (multinoulli) y la distribución multinomial. Así, la distribución Dirichlet puede ser utilizado como *prior* en un *finite mixture model* considerando $\pi \sim \text{Dir}(\frac{\alpha}{K} 1_K)$ y $\phi_k \sim H$.

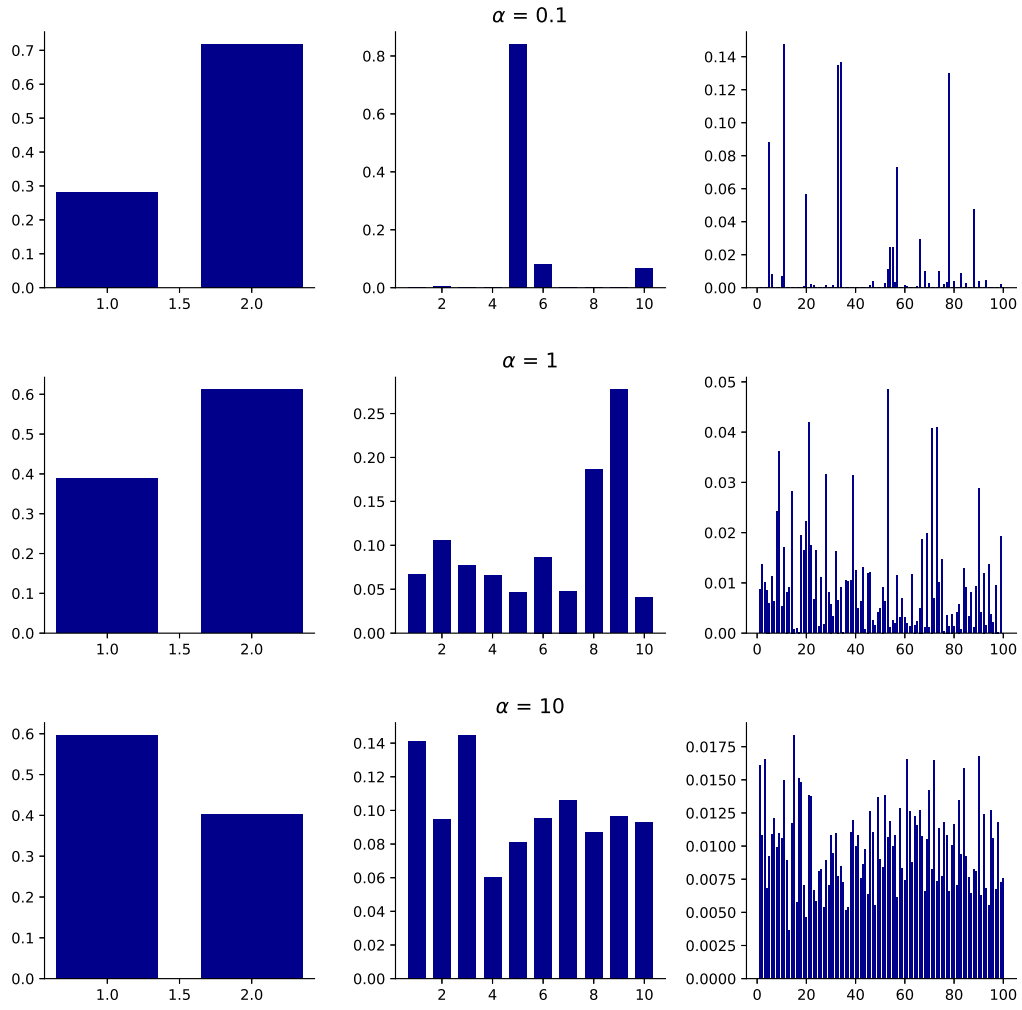


Figura 2.2: Muestra de una distribución Dirichlet simétrica con $\alpha \in \{0.1, 1, 10\}$ y $K \in \{2, 10, 100\}$.

2.1.2. Dirichlet Process

En un *finite mixture model* se tiene $G(\phi) = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}(\phi)$, luego al muestrear de G, con probabilidad uno se obtendrá exactamente K *clusters*. Nos gustaría tener un modelo más flexible, que pueda generar un número variable de *clusters*. Una forma de hacer esto es remplazar la distribución discreta G por una medida aleatoria de probabilidad, como el Dirichlet Process (Ferguson, 1973), denotado $G \sim \mathrm{DP}(\alpha, H)$.

Un **Dirichlet Process** (DP) es una distribución sobre medidas de probabilidad $G : \Phi \to \mathbb{R}^+$, donde $G(\phi) \geq 0$ y $\int_\Phi G(\phi) d\phi = 1$. Un DP se define implícitamente por cumplir

$$G(A_1), \ldots, G(A_K) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K)) \tag{2.11}$$

para cualquier partición finita $(A_1, \ldots, A_k)$ de Φ. En este caso, decimos que $G \sim \text{DP}(\alpha, H)$, donde α es llamado el **parámetro de concentración** y $H : \Phi \to \mathbb{R}^+$ es llamado la **medida base**.

Como $p(G(A_1), \ldots, G(A_K))$ es Dirichlet, la distribución marginal en cada partición distribuye beta $\text{Beta}(\alpha H(A_i), \alpha \sum_{j \neq i} H(A_j))$. El DP es considerado consistentemente definido, en el sentido de que si particionamos $\bar{A}_1$ en A_1 y A_2, entonces $G(\bar{A}_1)$ y $G(A_1) + G(A_2)$ siguen la misma distribución beta.

Sea $\phi \sim \text{Dir}(\alpha)$, y $z|\phi \sim \text{Cat}(\pi)$, si se integra π afuera se obtiene la distribución predictiva del modelo Dirichlet-multinoulli:

$$z \sim \text{Cat}(\alpha_1/\alpha_0, \ldots, \alpha_K/\alpha_0) \tag{2.12}$$

donde $\alpha_0 = \sum_k \alpha_k$. Es decir, $p(z = k|\alpha) = \alpha_k/\alpha_0$. Ademas, la posterior de π dada una observación viene dada por

$$\pi|z \sim \text{Dir}(\alpha_1 + \mathbb{I}(z = 1), \ldots, \alpha_K + \mathbb{I}(z = K)) \tag{2.13}$$

El DP generaliza el resultado anterior a particiones arbitrarias. Si $G \sim \text{DP}(\alpha, H)$, luego $p(\phi \in A_i) = H(A_i)$ y la posterior es

$$p(G(A_1), \ldots, G(A_K)|\phi, \alpha, H) = \text{Dir}(\alpha H(A_1) + \mathbb{I}(\phi \in A_1), \ldots, \alpha H(A_K) + \mathbb{I}(\phi \in A_K)) \tag{2.14}$$

Esto se mantiene para cualquier conjunto de particiones. Por lo tanto, si observamos multiples muestras $\bar{\phi}_{1:N} \sim G$, la nueva posterior está dada por

$$G|\bar{\phi}_{1:N}, \alpha, H \sim \text{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\phi_i}\right)\right) \tag{2.15}$$

Por ende el DP define un *prior* conjugado para cualquier espacio medible, donde el parámetro de concentración α es el tamaño de muestro efectivo de la medida base H.

Existen diferentes perspectivas que ayudan a entender la propiedad de *clustering* de un Dirichlet Process. En la sección 2.1.3. y 2.1.4. se describen dos: el Stick Breaking Process y Chinese Restaurant Process (CRP).

2.1.3. Stick Breaking Process

En esta sección se describe una definición constructiva de un DP, conocida como *stick breaking process* (Sethuraman, 1994). Sea $\pi = \{\pi_k\}_{k=1}^{\infty}$ una mezcla de pesos infinita derivada

a partir del siguiente proceso:

$$\beta_k \sim \text{Beta}(1, \alpha) \tag{2.16}$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l) = \beta_k(1 - \sum_{l=1}^{k-1}\pi_l) \tag{2.17}$$

Esto se suele denotar como $\pi \sim \text{GEM}(\alpha)$, donde GEM representa Griffiths, Engen y McCloskey, ver Figura 2.3 para una ilustración.

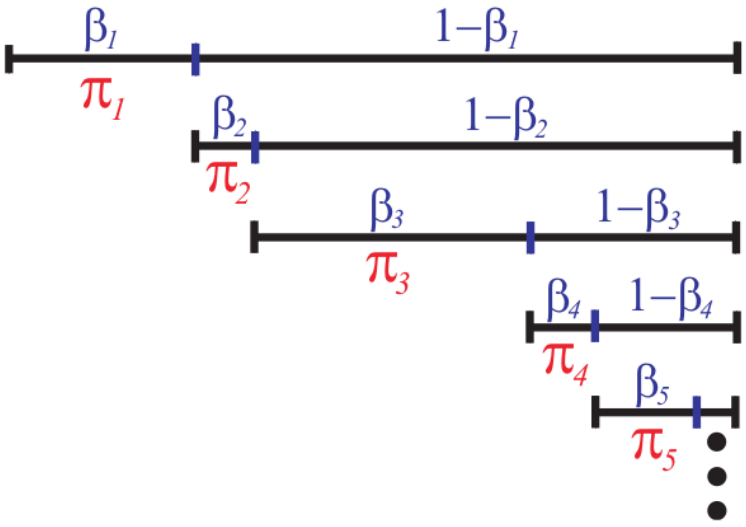


Figura 2.3: Ilustración de *stick breaking process*. Se tiene una barra de largo 1, la cual se rompe en un punto aleatorio β_1, el largo de la pieza restante es llamada π_1, luego recursivamente se rompe la barra restante, así generando $\pi_2, \pi_3, \ldots$. Fuente: Figura 2.22 de (Sudderth, 2006).

Algunos ejemplos de este proceso son mostrados en la Figura 2.4 (a). A mayor α, menos varianza y mayor número de átomos, por el contrario, pequeños valores de α muestran una alta varianza y menor número de átomos. Se puede demostrar que este proceso terminará con probabilidad uno, a pesar que el número de elementos que genera incrementa con α. Además, el tamaño del componente π_k decrece en promedio. La distribución G se puede definir como sigue:

$$G(\phi) = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(\phi) \tag{2.18}$$

, donde $\pi \sim \text{GEM}(\alpha)$ y $\phi_k \sim H$. Es posible demostrar que $G \sim \text{DP}(\alpha, H)$. Como consecuencia de esta construcción, las muestras de un DP son **discretas con probabilidad uno**. En otras palabras, al muestrear $\bar{\phi}_i \sim G$ se observarán valores repetidos, por lo que la mayoría de los datos vendrán de los ϕ_k con π_k más largos. En la Figura 2.4 (b) se muestra un par de medidas aleatorias generadas a partir de un DP con una medida base normal.

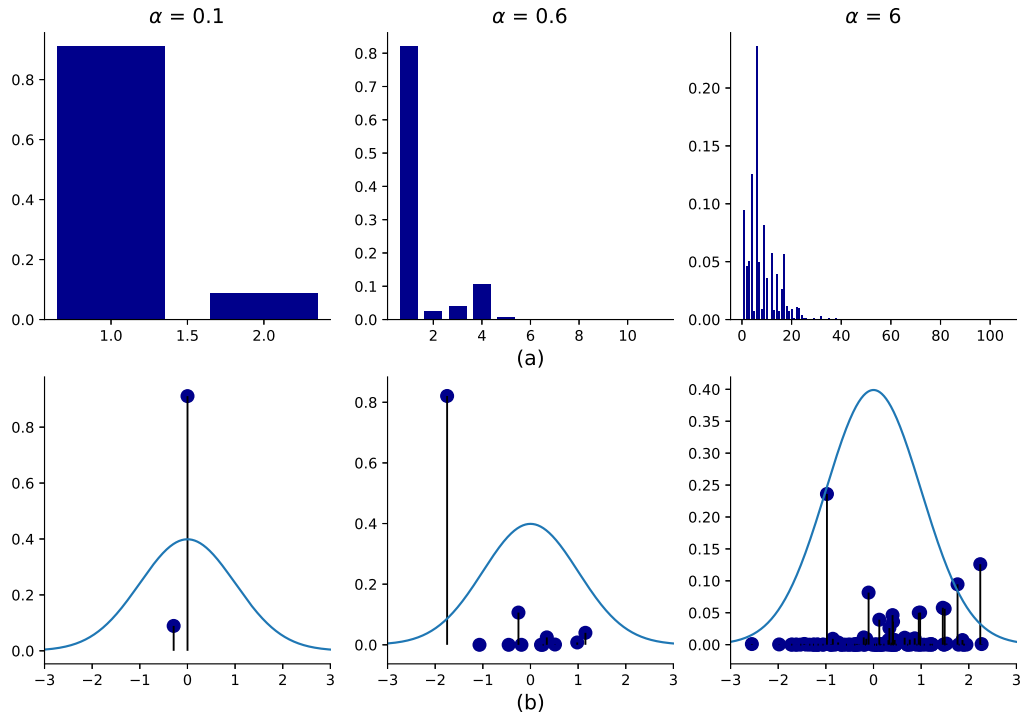


Figura 2.4: (a) Muestras de una distribución GEM con parámetros de concentración $\alpha \in \{0.1, 0.6, 6\}$. (b) Medidas aleatorias generadas a partir de un Dirichlet Process con medida base normal $\mathcal{N}(0, 1)$ con parámetros de concentración $\alpha \in \{0.1, 0.6, 6\}$

2.1.4. Chinese Restaurant Process

Trabajar con infinitos átomos puede ser bastante problemático. Para sortear esta dificultad se puede explotar la propiedad de *clustering* de un DP. Sea $\bar{\phi}_{1:N} \sim G$ observaciones generadas a partir de $G \sim \mathrm{DP}(\alpha, H)$, sea K los distintos valores de $\bar{\phi}_{1:N}$, luego la distribución predictiva condicionada en las N observaciones está dada por

$$p(\bar{\phi}_{N+1} = \phi | \bar{\phi}_{1:N}, \alpha, H) = \frac{1}{\alpha + N} \left(\alpha H(\phi) + \sum_{k=1}^{K} N_k \delta_{\bar{\phi}_k}(\phi) \right) \tag{2.19}$$

donde N_k es el número de observaciones previas iguales a ϕ_k. Este esquema de muestreo es llamado *Polya urn* o *Blackwell-MacQueen*.

Es más conveniente trabajar con variables discretas z_i que especifican cual ϕ_k usar, así, se define $\bar{\phi}_i = \phi_{z_i}$. En base a esta expresión se tiene lo siguiente:

$$p(z_{N+1} = z|z_{1:N}, \alpha) = \frac{1}{\alpha + N}\left(\alpha\mathbb{I}(z = k^*) + \sum_{k=1}^{K} N_k\mathbb{I}(z = k)\right) \tag{2.20}$$

, donde k^* representa un nuevo *cluster* que no ha sido usado aún. Este proceso es denominado Chinese Restaurant Process (CRP) (Aldous, 1985), basado en la oferta aparentemente infinita de mesas en ciertos restaurantes Chinos. La analogía es la siguiente: Las tablas del restaurante son los *clusters* y los clientes son las observaciones. Cuando una persona entra al restaurante, esta puede escoger sentarse en una tabla existente con probabilidad proporcional al número de personas ya sentadas en esa tabla (N_k), en otro caso, con una probabilidad decreciente a medida que más personas entran al restaurante (debido a $1/(\alpha + N)$) escogerá sentarse en una nueva tabla k^*. El resultado de este proceso es una distribución sobre particiones de los naturales, la cual es como una distribución de clientes a tablas.

El hecho de que las tablas actualmente ocupadas son más probables de obtener nuevos clientes se suele llamar el fenómeno del *rich get richer*. En efecto, se puede demostrar que la dsitribución del número de *clusters* que induce este *prior* es básicamente una ley de potencia, donde el número de tablas K con probabilidad 1 se aproxima a $\alpha log(N)$ cuando $N \to \infty$, mostrando que la complejidad del modelo crece logarítmicamente con el tamaño de los datos.

2.2. Modelos de tópicos

Los modelos de tópicos probabilísticos ayudan a descubrir los temas latentes (*clusters*) en una colección de documentos, como estos temas están conectados unos a otros y cómo cambian en el tiempo. Permiten resumir un gran colección de documentos a través de sus temas y organizarlos entorno a estos.

Los modelos probabilísticos tratan un tópico como una distribución de probabilidad discreta sobre el vocabulario del corpus, siendo un práctica habitual interpretar un tópico a partir de sus N palabras más probables. Por ejemplo, con $N = 5$ las palabras más probables de un tópico son: “llaves”, “domicilio”, “individuos”, “casa” y “porton”, por lo que una etiqueta valida para este tópico podría ser “portonazo”.

En *procesamiento del lenguaje natural* (NLP) se suele trabajar bajo la asumpción de **bag of words** (bolsa de palabras), es decir, tanto los documentos como las palabras son tratadas como intercambiables. Es importante hacer notar que itercambiabilidad no es equivalente a que las variables aleatorias son independientes e identicamente distribuidas. Más bien, intercambiabilidad esencialmente puede ser interpretado como condicionalmente independientes e identicamente distribuidas, donde el condicionamiento es con respecto a los parámetros de una distribución de probabilidad. Por lo tanto, el supuesto de intercambiabilidad es claramente un supuesto de simplificación cuya principal justificación es la construcción de algoritmos computacionales más eficientes.

Un *mixture model* que trabaja bajo la asumpción de *bag of words* es *mixture of unigrams* (Nigam et al., 2000), el cual asume que todos los documentos provienen de un solo *cluster*

dentro de un conjunto finito de K *clusters*. Los documentos de un *cluster* discuten solo un tópico particular z, y cada tópico z está asociado a una distribución categórica. Así, la verosimilitud de observar un documento d es

$$w|z \sim \text{Cat}(\theta_z) \tag{2.21}$$

$$p(w_1, \ldots, w_{N_d}) = \sum_{z=1}^{K} p(z) \prod_{i=1}^{N_d} p(w_i|z) \tag{2.22}$$

En las secciones 2.2.1-2.2.2 se describe en detalle dos modelos de tópicos probabilísticos, Latent Dirichlet Allocation (LDA) y Hierarchical Dirichlet Process (HDP), considerado la generalización no parámetrica de LDA, donde el número de tópicos a descubrir no está acotado y se infiere a partir del corpus.

En comparación a *mixture of unigrams*, LDA y HDP suponen que las palabras de un documento provienen de un mismo *mixture model*, donde a nivel corpus los *mixture models* comparten parámetros, que vienen siendo los tópicos, pero las *mixtures of topics* son específicas de cada documento. Esto permite relajar la asumpción de que cada documento es generado por un solo tópico, debido a que cada palabra proviene de algún tópico, por lo que un documento puede tener presencia de más de un tema.

2.2.1. Latent Dirichlet Allocation

En Latent Dirichlet Allocation (LDA) (Blei et al., 2003) cada tópico es una distribución de probabilidad sobre un vocabulario fijo V. Cada documento d tiene su propia mezcla de tópicos π_d. La asignación $z_{d,n} \in \{1, \ldots, K\}$ de de una palabra n a un tópico z es dibujada a partir de π_d. El modelo completo es como sigue

$$\phi_k|\eta \sim \text{Dir}(\frac{\eta}{|V|}1_{|V|}) \tag{2.23}$$

$$\pi_d|\alpha \sim \text{Dir}(\frac{\alpha}{K}1_K) \tag{2.24}$$

$$z_{d,n}|\pi_d \sim \text{Cat}(\pi_d) \tag{2.25}$$

$$w_{d,n}|z_{d,n}, \phi_{1:K} \sim \text{Cat}(\phi_{z_{d,n}}) \tag{2.26}$$

Esto es ilustrado en la Figura 2.5.

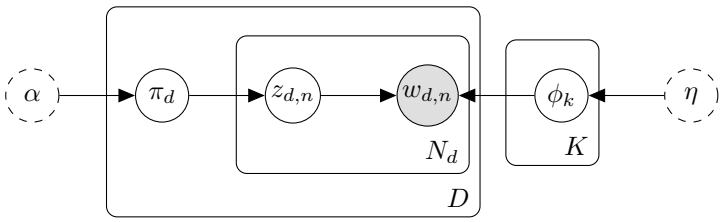


Figura 2.5: Representación gráfica de LDA: círculos denotan variables aleatorias, círculos abiertos denotan parámetros, círculos sombreados denotan variables observadas y los platos indican replicación.

La probabilidad conjunta del modelo:

$$p(\phi, \pi, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\phi_k|\eta) \prod_{d=1}^{D} p(\pi_d|\alpha) \prod_{n=1}^{N_d} p(z_{n,d}|\pi_d) p(w_{d,n}|\phi_{1:K}, z_{d,n}) \tag{2.27}$$

La distribución a posterior:

$$p(\phi, \pi, z|w, \alpha, \eta) = \frac{p(\phi, \pi, z, w|\alpha, \eta)}{p(w|\alpha, \eta)} \tag{2.28}$$

La distribución posterior es computacionalmente intratable para inferencia exacta, debido a que para normalizar la distribución se debe marginalizar sobre todas las variables ocultas y escribir la constante de normalización en términos de los parámetros del modelo. Para poder computar la posterior es necesario utilizar algoritmos de inferencia aproximada, donde el enfoque habitual es Markov Chain Monte Carlo (MCMC) (Andrieu et al., 2003) e Inferencia Variacional (VI) (Blei et al., 2017). En (Blei et al., 2003) se propone un algoritmo basado en VI y en (Griffiths and Steyvers, 2004) en MCMC.

Una representación equivalente en LDA sería generar cada palabra de un documento d a partir de un tópico dibujado por una distribución G_d,

$$\begin{aligned}
\phi_k|\eta &\sim \text{Dir}(\frac{\eta}{|V|} 1_{|V|}) && (2.29)\\
\pi_d|\alpha &\sim \text{Dir}(\frac{\alpha}{K} 1_K) && (2.30)\\
G_d(\phi) &= \sum_{k=1}^{K} \pi_{d,k} \delta_{\phi_k}(\phi) && (2.31)\\
\phi_{d,n}|\pi_d, \phi_{1:K} &\sim G_d && (2.32)\\
w_{d,n}|\phi_{d,n} &\sim \text{Cat}(\phi_{d,n}) && (2.33)
\end{aligned}$$

2.2.2. Hierarchical Dirichlet Process

Hierarchical Dirichlet Process (HDP)(Teh et al., 2005) es un *prior* jerárquico no paramétrico, el cual está formado por un DP cuya medida base G_0 es dibujada a partir de un DP. En el caso de modelamiento de tópicos, se tiene un medida global G_0 a nivel corpus que es dibujada a partir de un DP con medida base Dirichlet y una medida para cada documento

que es dibujada a partir de un DP cuya medida base es G_0. El modelo completo es como sigue

$$
\begin{aligned}
H &= \mathrm{Dir}(\frac{\eta}{|V|}1_{|V|}) && (2.34)\\
G_0|\gamma, H &\sim \mathrm{DP}(\gamma, H) && (2.35)\\
G_d|\alpha, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) && (2.36)\\
\phi_{d,n}|G_d &\sim G_d && (2.37)\\
w_{d,n}|\phi_{d,n} &\sim \mathrm{Cat}(\phi_{d,n}) && (2.38)
\end{aligned}
$$

Esto es ilustrado en la Figura 2.6.

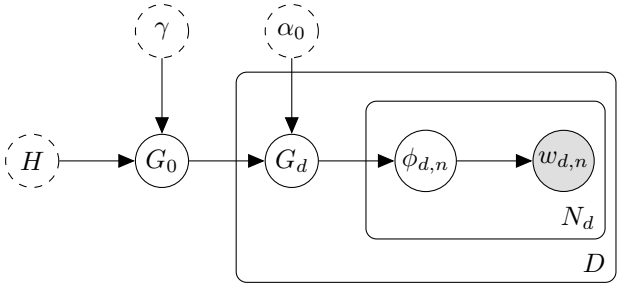


Figura 2.6: Representación gráfica de HDP: círculos denotan variables aleatorias, círculos abiertos denotan parámetros, círculos sombreados denotan variables observadas y los platos indican replicación.

La discretitud a nivel corpus de G_0 asegura que todos los documentos comparten el mismo conjunto de tópicos (*mixture components*). A nivel documento G_d hereda los tópicos de G_0, pero los pesos de cada tópico (*mixture proportions*) es específica del documento.

2.2.2.1. Stick Breaking Construction

Aplicando *stick breaking construction* se tiene que para el DP dibujado a nivel corpus la siguiente representación:

$$
\begin{aligned}
\beta_k^{'} &\sim \mathrm{Beta}(1, \gamma) && (2.39)\\
\beta_k &= \beta_k^{'} \prod_{l=1}^{k-1}(1 - \beta_l^{'}) && (2.40)\\
\phi_k &\sim H && (2.41)\\
G_0(\phi) &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}(\phi) && (2.42)
\end{aligned}
$$

Así, G_0 es discreto y tiene soporte en los átomos $\phi = \{\phi\}_{k=1}^{\infty}$ con pesos $\beta = \{\beta_k\}_{k=1}^{\infty}$, siendo la distribución de β escrita como $\beta \sim \mathrm{GEM}(\gamma)$. La construcción a nivel documento de G_d es:

$$\pi_{d,k}^{'} \sim \text{Beta}\Big(\alpha_0\beta_k, \alpha_0\Big(1-\sum_{l=1}^{k}\beta_l\Big)\Big) \tag{2.43}$$

$$\pi_{d,k} = \pi_{d,k}^{'}\prod_{l=1}^{k-1}(1-\pi_{d,l}^{'}) \tag{2.44}$$

$$G_d(\phi) = \sum_{k=1}^{\infty}\pi_{d,k}\delta_{\phi_k}(\phi) \tag{2.45}$$

$$\phi_{d,n}|\pi_d, \phi_{1:\infty} \sim G_d \tag{2.46}$$

Donde $\phi = \{\phi_k\}_{k=1}^{\infty}$ son los mismos átomos de G_0. Esto es ilustrado en la Figura 2.7.

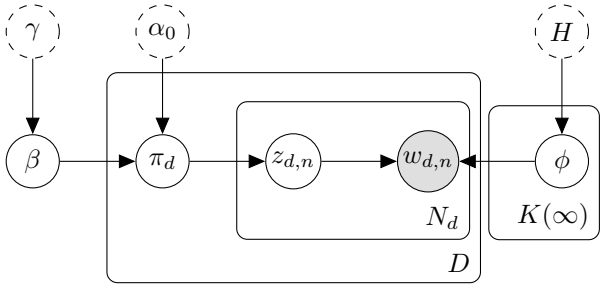


Figura 2.7: Representación gráfica de la construcción stick-breaking de HDP: círculos denotan variables aleatorias, círculos abiertos denotan parámetros, círculos sombreados denotan variables observadas y los platos indican replicación.

2.2.2.2. Chinese Restaurant Franchise Process

Una construcción alternativa de HDP es conocida bajo el nombre de *Chinese Restaurant Franchise Process* (CRF), una extensión del CRP, que permite compartir un conjunto de platos a través de una cadena de restaurantes Chinos. La analogía es la siguiente, se tienen D restaurantes, cada uno con N_d clientes que se sientan en tablas $t_{d,i}$, en cada tabla es servido un único plato $\phi_k \sim H$ a partir de un menú común para todos los restaurantes.

Sea m_{dk} el número de tablas sirviendo el plato k en el restaurante d, así $m_{d.}$ representa el número de tablas en el restaurante d, $m_{.k}$ representa el número de tablas sirviendo el plato k, y $m_{..}$ el número total de tablas ocupadas. Al integrar G_d la probabilidad condicional del cliente i-ésimo este en la tabla t se puede escribir como sigue:

$$p(t_{di}=t|t_{d1},\ldots,t_{d,i-1},\alpha_0,G_0) = \frac{1}{\alpha_0+i-1}\left(\alpha_0\mathbb{I}(t=t^*)+\sum_{t^{'}=1}^{m_{d.}}N_{dt^{'}}\mathbb{I}(t=t^{'})\right) \tag{2.47}$$

donde $N_{dt^{'}}$ representa los clientes del restaurante d que están sentados en la tabla $t^{'}$. Con probabilidad proporcional a los clientes sentados en la tabla t los clientes del restaurante se sentarán en esta y con probabilidad proporcional a α_0 en una nueva. Una vez todos los clientes estan sentados se tiene una partición sobre $\phi_{d1},\ldots,\phi_{dN_d}$ para cada documento d.

Luego, al integrar afuera G_0 se obtiene:

$$p(z_{dt} = z|z_{11}, z_{12}, \ldots, z_{d1}, \ldots, z_{d,t-1}|\gamma, H) = \frac{1}{\gamma + m_{\cdot\cdot}} \left(\gamma \mathbb{I}(z = k^*) + \sum_{k=1}^{K} m_{\cdot k} \mathbb{I}(z = k) \right) \quad (2.48)$$

en este caso se tiene que la tabla t del restaurante d con probabilidad proporcional al número de tablas que sirven el plato k ($m_{\cdot k}$) servirá el plato k y con probabilidad proporcional a γ servirá un nuevo plato.

Por último, al igual que LDA la distribución posterior de HDP es intratable, por lo que se debe recurrir a técnicas de inferencia aproximada. En (Teh et al., 2005) se propone un algoritmo basado en MCMC bajo la construcción CRF de un HDP.

2.3. Modelamiento de la evolución de los tópicos en el tiempo

En (Wilson and Robinson, 2011) y (Beykikhoshk et al., 2018) se propone una metodología que permite capturar los dinámismos mencionados usando LDA y HDP respectivamente. Donde se propone dividir el corpus en T épocas, en cada época se entrena un modelo de tópicos estático, obteniéndose así T conjuntos de tópicos $\phi = \{\phi_1, \ldots, \phi_T\}$, con $\phi_t = \{\phi_{t,1}, \ldots, \phi_{t,K_t}\}$ el conjunto de tópicos que describen la época t, y K_t el número de tópicos inferido en esa época. Una vez descubiertos los tópicos se hace uso de medidas de distancia o similitud para relacionar tópicos de épocas adyacentes.

En las secciones 2.3.1-2.3.2 se describe la metodología propuesta en (Beykikhoshk et al., 2018) para relacionar los tópicos descubiertos de épocas adyacentes.

2.3.1. Gráfo de similitud temporal

Para relacionar los tópicos de una época es necesario contar una medida de similitud $\rho \in [0, 1]$, con esta médida de similitud se puede construir un gráfo, donde los nodos son los tópicos de una época y los arcos relacionan tópicos de una época con la siguiente, siendo el peso del arco la similitud entre los tópicos. Una vez construido el grafo se eliminan las conexiones débiles en base a un umbral $\zeta \in [0, 1]$ a definir, reteniendo solo aquellas conexiones entre tópicos suficientemente similares entre épocas adyacentes, matemáticamente se poda el arco entre los tópicos $\phi_{t,i}$ y $\phi_{t+1,j}$ si $\rho(\phi_{t,i}, \phi_{t+1,j}) \leq \zeta$.

Está metodología permite detectar desaparición de un tópico, nacimiento de un nuevo tópico, como también división o fusión entre diferentes tópicos. A continuación se define en detalle cada uno de estos dinamismos:

- **Nacimiento de un tópico:** Si un tópico no tiene ningún arco entrante, por ejemplo, en la Figura 2.8 el tópico ϕ_{j+2} en t.
- **Muerte de un tópico:** Si un tópico no tiene ningún arco saliente, por ejemplo, en la Figura 2.8 el tópico ϕ_j en t.

- **Evolución de un tópico:** Cuando un tópico tiene exactamente un arco de entrada y salida, por ejemplo, en la Figura 2.8 entre las épocas t y $t+1$ se tiene que el tópico ϕ_{j+2} evoluciona del tópico ϕ_{k+1}.
- **División de un tópico:** Si un tópico tiene más de un arco saliente, por ejemplo, en la Figura 2.8 el tópico ϕ_i de $t-1$ se divide en $t+1$ en los tópicos ϕ_j y ϕ_{j+1}.
- **Fusión de un tópico:** Cuando un tópico tiene más de un arco entrante, este tipo de tópicos también pueden ser entendidos como un nuevo tópico, por ejemplo, en la Figura 2.8 los tópicos ϕ_i y ϕ_{i+1} de $t-1$ forman al tópico ϕ_{j+1} en t.

Una ilustración conceptual del grafo de similitud es mostrado en la Figura 2.8, este muestra tres épocas consecutivas.

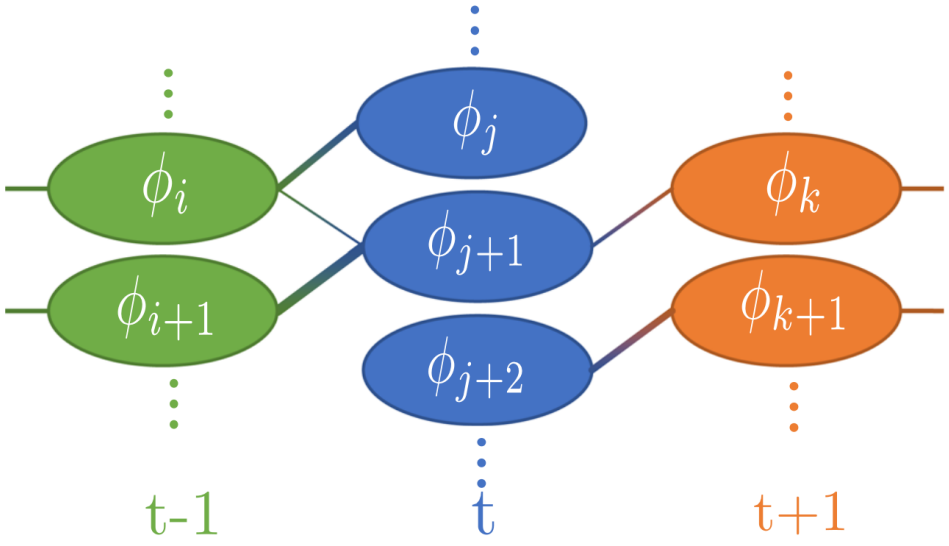


Figura 2.8: Ilustración conceptual del grafo de similitud que modela la dinámica de los tópicos en el tiempo. Un nodo corresponde a un tópico en una época específica; el ancho de los arcos es proporcional a la similitud entre los tópicos, arcos ausentes fueron eliminados por presentar una similitud menor a un umbral. Fuente: Figura 3 de (Beykikhoshk et al., 2018)

2.3.2. Construcción automática del grafo de similitud

Un aspecto relevante de esta metodología es definir el úmbral de corte, el cual no es fácilmente interpretable, además el úmbral depende de la médida de similitud escogida, dificultando así la comparación entre médidas de similitud. En (Beykikhoshk et al., 2018) proponen una alternativa más interpretable para definir el úmbral, para esto estiman la función de densidad acumulada (cdf) del grafo inicial, donde todos los nodos de una época están conectados con todos los nodos de la época adyacente, al que llamaremos grafo *fully connected*.

Sea F_p la cdf sobre las similitudes del grafo inicial, luego sea $\zeta \in [0,1]$ el punto operante de la cdf, luego eliminamos el arco entre los tópicos $\phi_{t,i}$ y $\phi_{t+1,j}$ si $\rho(\phi_{t,i}, \phi_{t+1,j}) \leq F_p^{-1}(\zeta)$, donde

$F_p^{-1}(\zeta)$ es el cuantil ζ de F_p. En 2.9 se tiene una ilustración para tres médidas de similitud, en esta se observa que la elección de un úmbral de corte arbitrario depende fuertemente de la médida de similitud escógida, por lo que la elección en base a la cdf puede ser más apropiada.

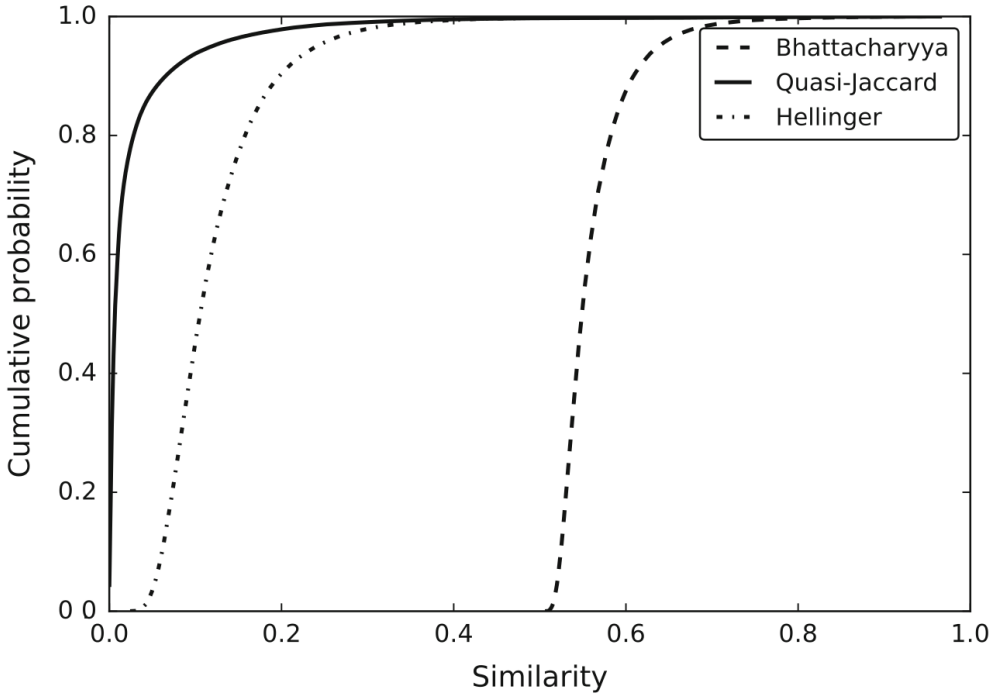


Figura 2.9: Estimación empírica de la función de densidad acumulada (cdf) de la similitud entre tópicos de épocas adyacentes en un grafo *fully connected* para tres medidas de similitud. Fuente: Figura 4 (Beykikhoshk et al., 2018).

Capítulo 3

Metodología

En este capítulo se describe la metodología propuesta para el descubrimiento de tópicos y su evolución en el tiempo. En primer lugar, en la sección 3.1 se describe la metodología de procesamiento utilizada para limpiar los datos que usará el modelo. En segundo lugar, en la sección 3.2 se justifica la elección del modelo de tópicos junto a la herramienta usada para facilitar la interpretación de estos. En tercer lugar, en la sección 3.3 se describe la metodología escogida para modelar la evolución en el tiempo y la médida de similitud utilizada para comparar tópicos de épocas adyacentes. En cuarto lugar, en la sección 3.4 se describen los hiperparámetros y su configuración. Por último, en la sección 3.5 se resume de manera global la metodología.

3.1. Procesamiento

El propósito del procesamiento en NLP es simplificar los datos lo más posible tal que se mantiene el *core* de palabras del corpus. En el caso del modelamiento de tópicos, esta etapa puede reducir significativamente el vocabulario. Como consecuencia, esto puede traer una mejora en la significancia estadística de los modelos, puesto que se puede obtener un mejor balance entre cantidad de parámetros y observaciones. Adicionalmente, puede facilitar la interpretación de los tópicos, removiendo palabras que aportan poca información.

En este experimento se aplicaron las siguientes cinco etapas:

1. **Tokenización**: La tokenización es una operación sobre una cadena de caracteres (*string*) que consiste en dividir el *string* en un conjunto de términos (ej: en base al caracter espacio), obteniéndose así una lista de elementos llamados *tokens*, que en términos simples pueden considerarse como palabras.

2. **Procesamiento de caracteres**: En esta etapa se suelen aplicar algunas operaciones básicas de procesamiento. En este proceso se llevan los tokens a unicode y minúsculas. Luego, se eliminan patrones de caracteres que difícilmente pueden tener algún significado, como correos electrónicos, símbolos de puntuación, tokens con números y letras o solo números.

3. **Eliminación de stopwords**: Las *stopwords* (Wilbur and Sirotkin, 1992) son palabras que aportan poca información (ej: artículos, preposiciones y conectores), usualmente tienen un alta frecuencia dentro del corpus. Para esto se utilizó una lista de palabras

disponibles en el paquete NLTK de Python, el cual contiene 313 palabras (Bird et al., 2009). Además, esta lista es alimentada con 951 *stopwords* contextuales que corresponden a palabras específicas del corpus que aportan poca información. En el caso del robo de vehículos palabras relacionas a "robo" o "vehículo" no aportan ninguna información, puesto a que todos los relatos hablan del robo de un vehículo.

4. **Filtro por vocabulario**: Con el propósito de mantener palabras "humanamente legibles" se utiliza un vocabulario. Para esto se utilizó el vocabulario del corpus SUC descrito en el capítulo anterior, de esta menera toda palabra tiene su *word embedding*.

5. **Filtro por frecuencia**: En este nivel se eliminan los *tokens* con baja frecuencia. Esta etapa viene motivada del hecho de que un modelo difícilmente aprenderá algún patrón de un evento que tiene muy pocas realizaciones, menos si tiene una realización única. Esta etapa se aplicó a nivel época, eliminando aquellos tokens que aparecen en menos del 0.1 % de los documentos de su respectiva época.

6. **Eliminación de documentos**: En el último nivel se eliminan aquellos documentos que presentan pocos tokens. Esto tiene por objetivo obtener estimaciones más confiables y reducir la posibilidad de sacar conclusiones prematuras debido a las pocas observaciones con las que cuenta un documento. En este caso se elimnarón documentos que presentan menos de 5 tokens.

3.2. Modelos de tópicos

Se escoge HDP como el modelo de tópicos base de la metodología. Si bien, HDP es un modelo similar en estructura a LDA, su principal ventaja es que el número de tópicos no está acotado y es inferido a partir de los datos, en cambio LDA requiere de escoger el número de tópicos K por adelantado.

En un enfoque tradicional, se requiere de entrenar múltiples veces LDA para diferentes valores de K y se escoge la configuración con mejor desempeño en un conjunto de validación, por lo que LDA termina siendo computacionalmente más costoso que HDP, además este enfoque se vuelve impracticable cuando el conjunto de datos es lo suficientemente grande.

En el aspecto cualitativo ambos modelos entregan tópicos igual de consistentes. En cuanto a métricas de desempeño como *perplexity* HDP suele tener mejor desempeño (Teh et al., 2005).

Para el descubrimiento de tópicos se utilizó la implementación disponible en C++ (Wang and Blei, 2010) de HDP para modelamiento de tópicos. Esta implementación está basada en el algoritmo de Gibbs Sampling propuesto en (Teh et al., 2005).

3.2.1. Interpretación de tópicos

Los modelos de tópicos probabilísticos se caracterizan por tener un alto poder interpretativo, esto se debe a que la distribución de probabilidad de cada tópico sobre el vocabulario da una idea del tema al que pertenece, por otro lado, la mezcla de tópicos de cada documento muestra que tan importante es cada tópico en la generación de estos, como también dentro

del corpus.

En este sentido, las visualizaciones pueden ayudar a interpretar mejor los resultados de los modelos de tópicos. Para la interpretación de los tópicos la metodología propuesta se basa en la herramienta de visualización desarrollada en (Sievert and Shirley, 2014), la cual responde las siguientes preguntas, ¿Cuál es el significado de cada tópico?¿Cuán predominante es cada tópico?¿Cómo se relacionan los tópicos entre sí?

Para responder la pregunta 1 se incorpora un gráfico de barras que muestra las palabras más relevantes del tópico seleccionado dado un parámetro $\lambda \in [0, 1]$. A través de una visualización espacial responde la pregunta 2 y 3. La visualización espacial consiste en aplicar técnicas de reducción de dimensionalidad como TSNE (Maaten and Hinton, 2008) o PCA (Wold et al., 1987) a la matriz de distancia entre tópicos, usando Jensen-Shannon divergence (Endres and Schindelin, 2003) como médida de distancia. Una vez cada tópico es mapeado a un punto en un espacio de dos dimensiones se dibuja un círculo con centro en este punto y con radio proporcional a la cantidad de tokens generados por el tópico.

Para interpretar un tópico, lo usual es examinar una lista ordenada de las palabras más probables del tópico, usando ya sea desde cinco a treinta términos. Un problema frecuente que se presenta en este caso es que los términos que son comunes al corpus frecuentemente aparecen en el top de las palabras más probables de un tópico, haciendo difícil discernir el significado de estos. Para esto en (Sievert and Shirley, 2014) se define una métrica denominada *relevance*, la cual define la relevancia de una palabra no solo por su probabilidad dentro del tópico sino también por su exclusivad dentro del corpus. La *relevance* de una palabra w en el tópico k dado λ está dada a través de la siguiente expresión:

$$r(w, k|\lambda) = \lambda log(\phi_{kw}) + (1 - \lambda)\lambda log\left(\frac{\phi_{kw}}{p_w}\right) \tag{3.1}$$

, donde λ determina el peso que se le da a la probabilidad de la palabra w dentro del tópico k (ϕ_{kw}) relativo a su *lift*, el cual se define por el ratio entre la probabilidad de la palabra dentro del tópico y su probabilidad marginal a lo largo del corpus (p_w). Fijando $\lambda = 1$ se obtiene el ranking de términos decrecientes en orden de su probabilidad dentro del tópico, y fijando $\lambda = 0$ el ranking se basa solo en el *lift*.

3.3. Construcción del grafo temporal

El objetivo del trabajo no es solo descubrir tópicos sino también modelar sus interacciones en el tiempo, como nacimiento, muerte, evolución, divisón y fusión. Así, la metodología propuesta se basa en la metodología descrita en la sección 2.3.1, debido que esta captura los dinamismos mencionados.

En general, las medidas de similitud o distancia comparan vectores con el mismo dominio y dimensión, esto significa que los tópicos de épocas adyacentes deben compartir el mismo vocabulario. Matemáticamente, sea $\phi_{t,i}$ un tópico de la época t y V_t su vocabulario, sea $\phi_{t+1,j}$ un tópico de la época $t+1$ y V_{t+1} su vocabulario. Con una alta probabilidad existen palabras

en V_t que no están en V_{t+1} y viceversa. Para poder comparar tópicos en épocas adyacentes es necesario contar con un vocabulario global $V'_{t+1} = V_t \cup V_{t+1}$, luego aplicar *padding* a los vectores $\phi_{t,i}$ y $\phi_{t+1,j}$, es decir, rellenar con ceros las posiciones que no están en el vocabulario de su dominio.

Una gran desventaja del enfoque anterior es que no captura similitud entre palabras, puesto que cada palabra ocupa una posición dentro del vector y no hay forma de comparar palabras que no son comúnes en ambas épocas. El peor caso sería considerar los vocabularios V_t y V_{t+1}, con $V_t \cap V_{t+1} = \emptyset$, a pesar de que cada palabra en V_t tiene un sinónimo en V_{t+1} la similitud entre tópicos entre las épocas t y $t+1$ sería cero.

3.3.1. Word Mover's Distance

Para lidiar con el problema anterior, se propone utilizar una medida de distancia conocida como Word Mover's Distance (WMD) (Kusner et al., 2015), medida utilizada para comparar dos documento bajo una representación *bag of words* a través de sus *word embeddings* (Mikolov et al., 2013).

WMD calcula el costo mínimo de transformar un documento en otro, en esto caso particular sería el costo mínimo de llevar un tópico a otro. Para esto se resuelve el problema de transporte, donde los flujos son los pesos $\phi_{t,i}$ y $\phi_{t+1,j}$ y la matriz de costos es una matriz de distancia euclidiana entre los *word embedding* de todas las palabras de V_t con V_{t+1}. En la Figura 3.1 se ilustra el espacio en el que viven las palabras de dos documentos.

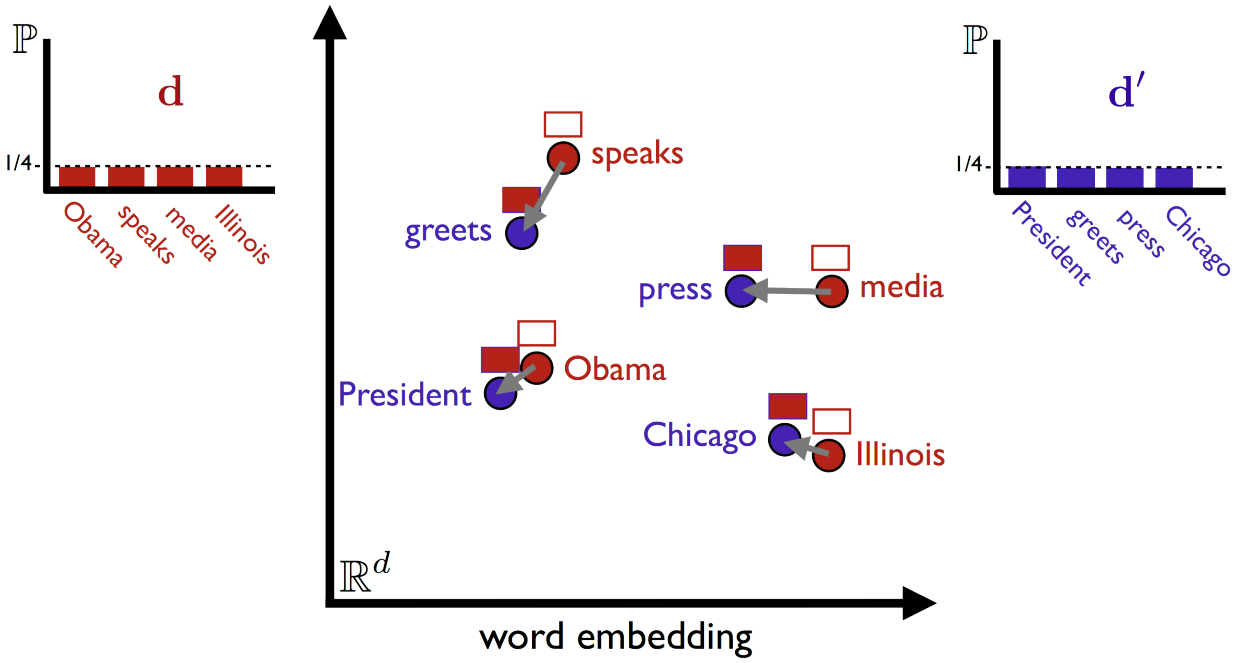


Figura 3.1: Espacio vectorial de los *word embeddings* de las palabras de dos documentos con un vocabulario de tamaño 4. Fuente: Figura de (Niculae, 2015).

Sea V_i y V_j los vocabularios del tópico i y j respectivamente, luego su WMD viene dado por $WMD(\phi_i, \phi_j)$:

$$\min_x \sum_{u \in V_i} \sum_{v \in V_j} c_{u,v} x_{u,v} \tag{3.2}$$

$$\text{s.t.} \sum_{v \in V_j} x_{u,v} = \phi_{i,u},\ u \in V_i \tag{3.3}$$

$$\sum_{u \in V_i} x_{u,v} = \phi_{j,v},\ v \in V_j \tag{3.4}$$

$$x_{u,v} \geq 0,\ u \in V_i\ , v \in V_j \tag{3.5}$$

Donde $x_{u,v}$ es el flujo que va de la palabra u del tópico i a la palabra v del tópico j, $\phi_{i,u}$ es la probabilidad de la palabra u en el tópico i, $c_{u,v}$ es el costo de mover una unidad de flujo por el arco (u, v), el costo entre palabras se mide como la distancia euclidiana entre los *word embedding* de dichas palabras.

La primera restricción indica que el flujo que se mueve de una palabra u del tópico i a todas las palabras del tópico j debe sumar su peso ($\phi_{i,u}$), la segunda restricción significa que el flujo que se mueve de una palabra v del tópico j a todas las palabras del tópico i debe sumar su peso ($\phi_{j,v}$). Lo anterior implica que esta medida de distancia es simétrica, es decir, $WMD(\phi_i, \phi_j) = WMD(\phi_j, \phi_i)$.

La WMD se puede transformar fácilmente en una médida de similitud considerando $\rho(\phi_i, \phi_j) = \frac{1}{1+WMD(\phi_i,\phi_j)}$. Notar que si la WMD es 0 la similitud es 1 y si es ∞ la similitud es 0.

3.3.2. WMD complejidad

WMD es una medida de distancia intensiva en recursos computacionales. Para mejorar el entendimiento se utiliza la representación poliedral del problema. Sea N el tamaño del vocabulario entre dos épocas adyacentes, luego la región factible del problema anterior se puede representar como $\{x|Ax = b, x \geq 0\}$ sobre un grafo bipartito, con $A \in \mathbb{R}^{2N \times N^2}$ la matriz de incidencia, $b \in \mathbb{R}^{2N}$ la capacidad de los nodos y $x \in \mathbb{R}^N$ el flujo a enviar por cada uno de los arcos. Para resolver este problema se utilizó la implementación de (Doran, 2014), la cual está basada en el algoritmo (Pele and Werman, 2009), cuya complejidad del mejor tiempo promedio escala $\mathcal{O}(N^2 logN)$.

Los tópicos siguen una distribución con forma de ley de potencia sobre el vocabulario, donde una pequeña fracción de las palabras concentran la mayor parte de la masa de la distribución. Además, en la práctica la interpretación de los tópicos se basa en los top N palabras más probables, usualmente con $N \in [5, 30]$, entonces, se puede aprovechar esta estructura para efectos de computar la WMD de un forma más eficiente, por ejemplo, utilizando solo las palabras que capturan un determinado porcentaje de la distribución acumulada del tópico. Por ejemplo, si se reduce el vocabulario a un décimo en el peor caso promedio se obtiene un

speed up de 200.

3.3.3. Word Embeddings

Computar WMD requiere contar con *word embeddings*. Para estó se utilizó una de las más grandes colecciones de *word embeddings* en español (Cañete, 2019a), que cuenta con 1,313,423 *embeddings*, colección obtenida utilizando el algoritmo FasText (Bojanowski et al., 2017) sobre el corpus Spanish Unannotated Corpora (SUC) (Cañete, 2019b), uno de los más grandes corpus de texto en español. FasText en comparación a otros enfoques para extraer *embeddings* representa los *tokens* a través de n-gramas de caracteres, de esta manera puede obtener *embeddings* de *tokens* no vistos durante el entrenamiento a partir de los *embeddings* de los caracteres que lo componen.

3.4. Configuración de hiperparámetros

En la metodología propuesta se pueden considerar dos fuentes de evaluación de desempeño, el descubrimiento de tópicos y cómo se relacionan. En ambos casos no se cuenta con el *ground truth* para medir correctamente el desempeño. Si se conociera el *ground truth*, se podría utilizar la métrica *purity* (Manning et al., 2008) para comparar la asignación de los documentos en torno a los tópicos con la etiqueta. En el caso del grafo temporal, si se conocieran las conexiones presentes y ausentes se podrían utilizar métricas de clasificación.

3.4.1. Configuración de hiperparámetros de HDP

HDP cuenta con tres hiperparámertos, el parámetro de concentración a nivel corpus γ, el parámetro de concentración a nivel documento α_0 y η el parámetro de la medida base Dirichlet.

En (Blei et al., 2003; Griffiths and Steyvers, 2004; Cao et al., 2009; Arun et al., 2010; Deveaud et al., 2014; Zhang et al., 2017) se describen algunas métricas que no requieren de una etiqueta, que pueden ser útil para realizar selección de modelo de tópico. Cabe destacar que estas métricas carecen de significado y sirven para comparar si un modelo de tópicos es superior a otro.

En general, en modelamiento de tópicos se prefiere usar $\eta \in (0, 1)$, esto generará distribuciones *sparse* sobre el vocabulario. Así, se suelen tener tópicos más distinguibles, donde el *core* de palabras del tópico concentra la masa de la distribución. Además, como la semántica del tópico está compactada en pocas palabras se facilita la interpretación. En este caso se utilizó un punto intermedio, fijando $\eta = 0.5$.

En (Teh et al., 2005) los parámetros de concentración se integran afuera usando un prior *vague gamma* (Escobar and West, 1995). Un prior *vague gamma* es una distribución Gamma con una gran parte de la masa en torno a cero y una cola pesada. Véase la Figura 3.2 para una ilustración de la pdf para diferentes parámetros. Por consecuencia, el prior tendrá un

menor efecto de regularización y a medida que más datos se obtienen la posterior coincidirá con las observaciones empíricas. En este caso se utilizó un prior $\Gamma(\alpha = 1, \beta = 1)$.

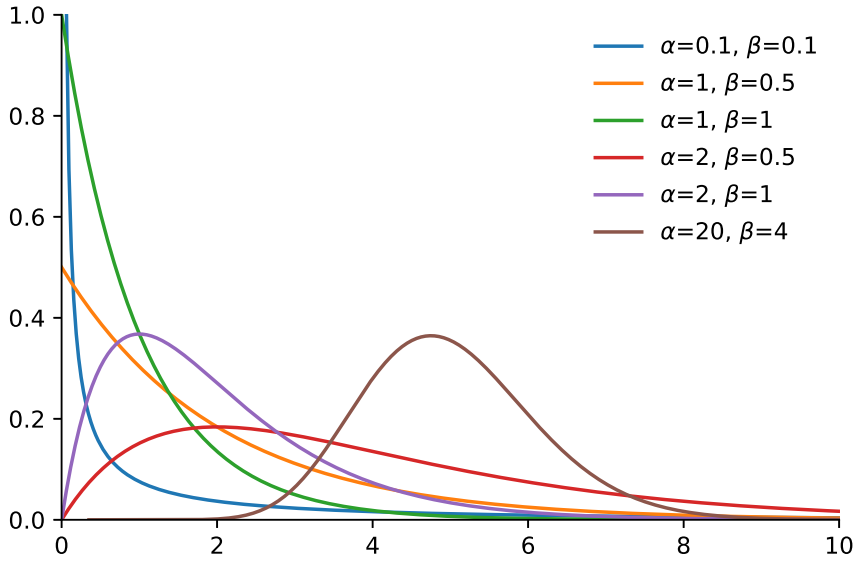


Figura 3.2: Función de densidad de probabilidad (pdf) de una distribución Gamma para diferentes parámetros de forma α y tasa β.

3.4.2. Configuración de hiperparámetros del grafo temporal

El grafo temporal cuenta con dos hiperparámetros: q y ζ. El parámetro $q \in [0, 1]$ define el soporte de los tópicos, utilizando aquellas palabras más probables que explican 100q % de la distribución acumulada del tópico. Por otro lado, el hiperparámetro $\zeta \in [0, 1]$ define el úmbral de corte, representa el punto operante de la cdf del grafo *fully connected*, permite definir el cuantil que se usará como úmbral para eliminar arcos con similitud menor a este.

En cuanto al parámetro q un valor razonable podría estar entre $[0.8, 0.95]$, de esta manera se conserva el *core* de palabras del tópico y se disminuye de manera significativa el tiempo de cómputo. Por otro lado, valores razonables de ζ podrían estar entre $[0.9, 0.99]$, de esta manera solo se conservarían aquellas relaciones con una alta similitud relativa, debido a que el umbral de corte no depende de la medida de similitud utilizada.

3.5. Resumen metodología

En la Figura 3.3 se presenta un esquema que resume de la metodología propuesta para el descubrimiento y seguimiento de tópicos en el tiempo. En primer lugar, se divide el corpus original en épocas y a cada época se le aplican las siguientes cinco etapas en forma secuencial: tokenización, procesamiento de caracteres, eliminación de *stopwords*, filtro por vocabulario y filtro por frecuencia. Como producto del proceso anterior se obtiene un corpus listo para la aplicación de un modelo de tópicos. En segundo lugar, se aplica HDP de forma independiente

sobre cada una de las épocas. Por último, una vez descubiertos los tópicos se procede a construir el grafo temporal, para esto es necesario computar la WMD entre los tópicos de épocas adyacentes. Finalmente, se podan los arcos cuya similitud es menor al cuantil ζ de la distribución acumulada del grafo *fully connected*.

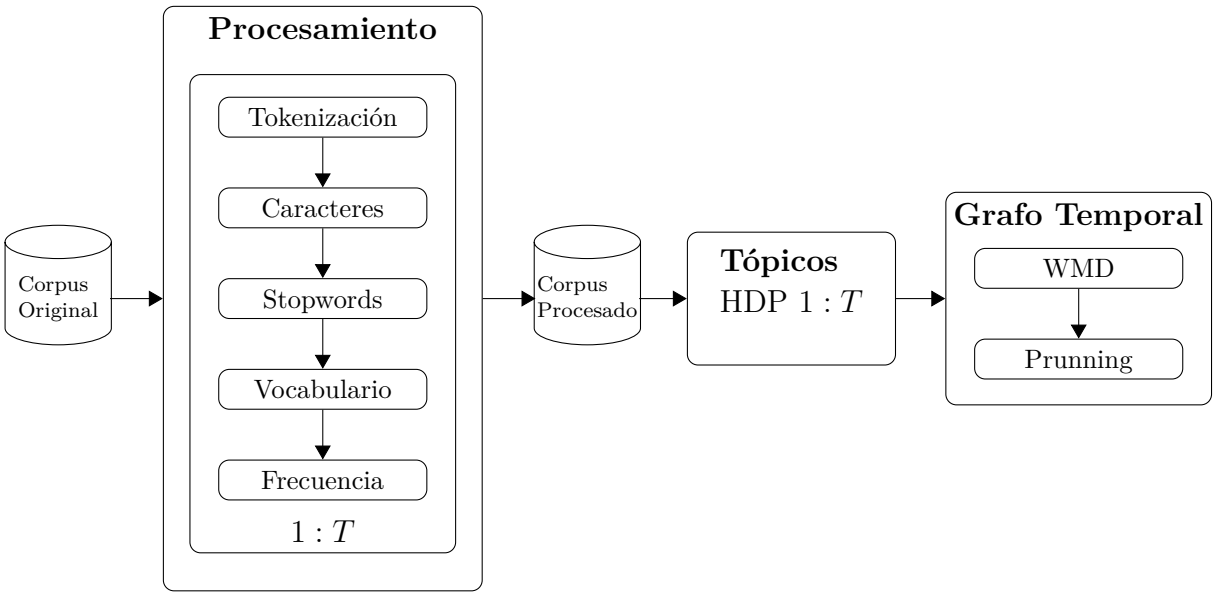


Figura 3.3: Esquema de la metodología de descubrimiento y evolución de tópicos.

Capítulo 4

Caso de estudio

En este capítulo se analizan los resultados de la metodología propuesta al fenómeno de robo de vehículos. En la sección 4.1 se describen la fuente de información utilizada. En la sección 4.2 los resultados del procesamiento de los datos. Por último, en la sección 4.3-4.4 se realiza un análisis cuantitativo y cualitativo de los resultados.

4.1. Datos

Para este experimento se cuenta con relatos de víctimas del robo de vehículos provistos por la Asociación de Aseguradores de Chile (AACH). Esta base de datos consta con 49,015 relatos entre los años 2011-2016, veasé la Figura 4.1 para el detalle por año.

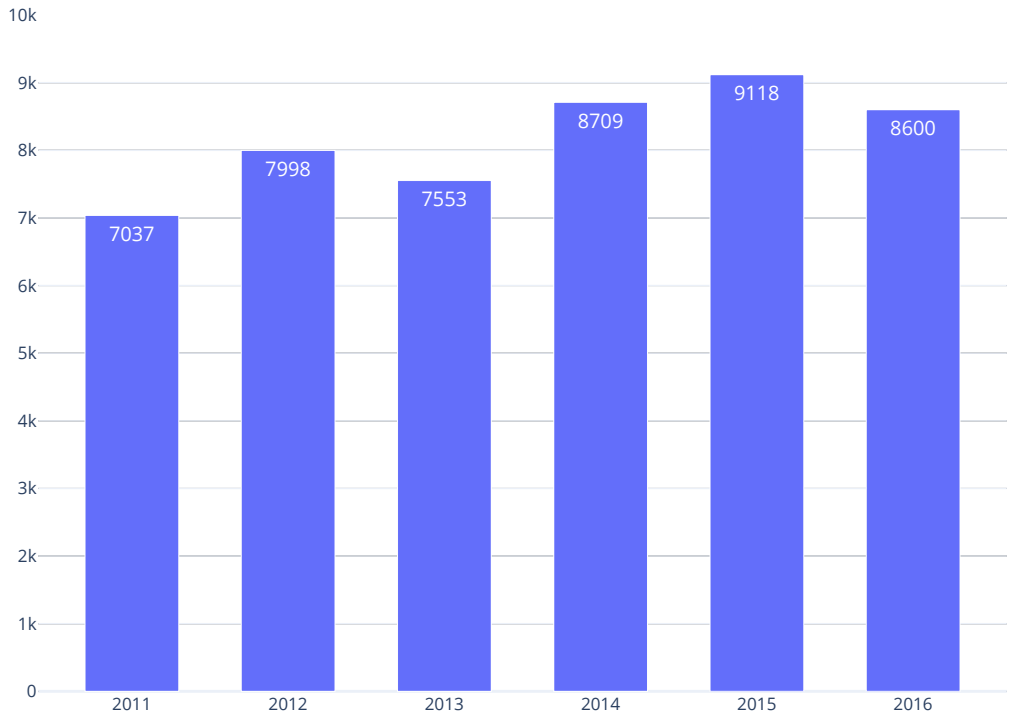


Figura 4.1: Cantidad de robos registrados por año en base de datos AACH.

En la Figura 4.2 se muestra algunos ejemplos de la base de datos de la AACH, de aquí se observa que los relatos carecen de estandarización y presentan múltiples errores ortográficos. En consecuencia, la etapa de procesamiento toma suma relevancia, ya que aplicar un modelo de tópicos a un corpus sin ningún tipo de procesamiento nos puede llevar a resultados no deseados. En la sección 4.2 se detallan los resultados obtenidos sobre el corpus tras aplicar los niveles de procesamiento mencionados en la sección 3.1.

ESTABA ESTACIONADO EL LA CALLE ROTEMBURGO ENTRE NORUGA Y SEÑORA DEL ROSAIO Y AL MOMENTO DE IR A BUSCAR EL AUTO SE DA CUENTA QUE EL VH NO SE ENCUETRA AL PARECER LO ROBARON. NO POSEEE LOS DOCUMENTOS DEL VH vh aparece pero con mulples daños e evaluar queda en manos del liquidador.

ME ENCONTRABA CARGANDO COMBUSTIBLE EN LA SHELL DE CARRASCAL CON WALKER MARTÍNEZ Y REPENTINAMENTE FUI ASALTADA EN FORMA VIOLENTA LLEVASE MI VEH (TENGO GRABACIÓN). DAÑOS: ROBO DE MI VEH . LEIVA SE DERIVA A DON MARIO MEDINA .3 UF.DED/XX@XX.CL

TEXT : DEJO MI VEHICULO ESTACIONADO EN DICHO LUGAR AL VOLVER ME PERCATO QUE EL VEHICULO HABIA SIDO ROBADO EL MISMO DIA DEL ROBO A LAS 20:00 SOY CONTACTADO POR CARABINEROS DE LA COMUNA DE EL BOSQUE LOS CUALES ME INFORMAN QUE HABIAN RECUPERADO MI VEHICULO EL CUAL PRESENTABA LOS SIGUIENTES DAÑOS : VIDRIO TRASERO DERECHO QUEBRADA CHAPA DE CONTACTO FORZADA PARACHOQUE DELANTERO DERECHO RAYADO ALARMADESCONECTADA OTROS DAÑOS EN EL SISTEMA ELECTRICO ALARMA DE AIRBAGS ENCENDIDA ROBO DE ESPECIES.

Descripción Siniestro: el dia 24 de abril se le arrendo el vh a XX el cual estuvo sin problemas pagando el arriendo hasta el mes pasado que no pago mas y se le ha llamado en reiteradas veces y dice que va a venir a dejar el auto y no aparecel. por eso se realizo una denuncia por apropiacion indevida

ammg 53966748 vh asegurado transitaba en calle copiapo alt. 750 en este punto sufro portonazo sujetos armados roban mi vh hoy a las 04.30am vh fue encontrado en sector de la pintana mi vh ahora esta siendo periciado. daños por evaluar

PATENTE XX Siendo las 22:30 en la interseccion de san Alfonso con Claudio Gay un individuo me obliga a bajar del vehiculo apuntandome con una pistola de inmediato aparecen dos personas mas las que me suben en la parte trasera del furgon donde constantemente me amenazan con dispararme me bajan del vehiculo en un potrero cercano a la autopista del sol teniendome boca abajo golpeandome luego me colocan un pa?o en la cara perdiendo el conocimiento al despertar desorientado me dirijo a car

Figura 4.2: Muestra de relatos de la base de datos AACH.

4.2. Procesamiento

En esta sección se detallan los resultados de aplicar el procesamiento descrito en la sección 3.1. Con fines gráficos los resultados del procesamiento se decriben en un orden distinto al descrito en dicha sección, con el objetivo de mostrar en como estas afectan el tamaño del vocabulario. El orden es el siguiente, (i) tokenización, (ii) procesamiento de caracteres, (iii) eliminación de palabras poco frecuentes, (iv) filtro por vocabulario, (v) eliminación de *stopwords* y (vi) eliminación de documentos con pocas palabras.

En la Figura 4.3 se muestran la distribución acumulada del corpus original tras solo aplicar tokenización. En este caso los *tokens* totales corresponden a 2,030,980 asociado a un vocabulario de 93,203 palabras.

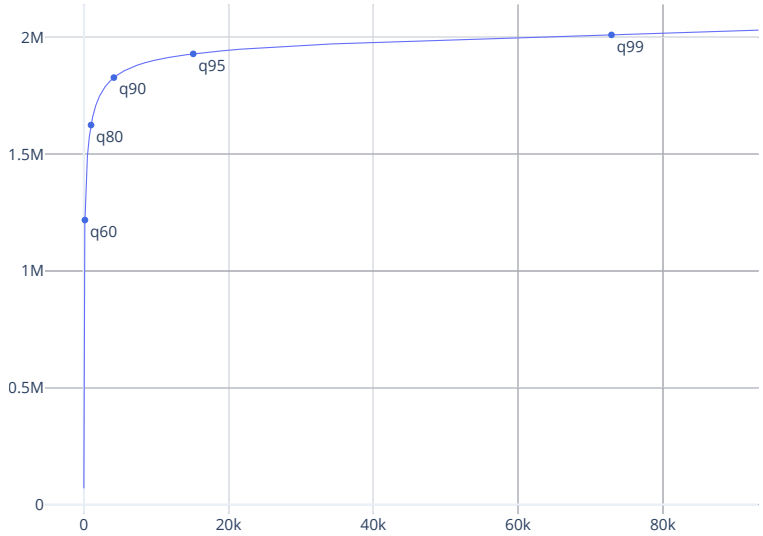


Figura 4.3: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el primer nivel de procesamiento.

La Figura 4.4 muestra los resultados al aplicar la etapa de procesamiento de caracteres, de esta se observa que se reduce el tamaño del vocabulario en cerca de la mitad, específicamente a 42,921 palabras, similarmente con la cantidad de *tokens*, que ahora son 1,028,412.

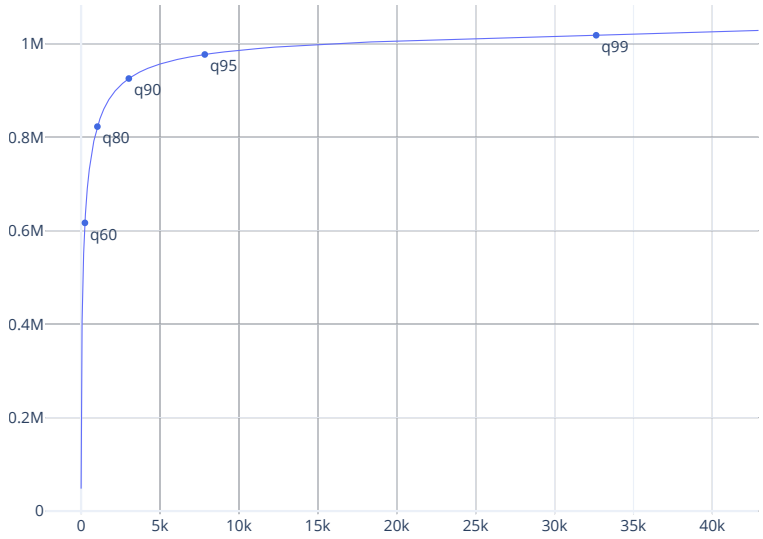


Figura 4.4: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el segundo nivel de procesamiento.

Hasta este nivel de procesamiento se tiene que al menos el 50 % de las palabras ocurren una única vez y al menos un 80 % tiene una frecuencia igual o menor a 4. El 95 % de la distribución acumulada puede ser explicada con 7,837 palabras (un 18 % del vocabulario actual). En conclusión, la distribución de las palabras tiene una cola bastante pesada.

En la Figura 4.5 se muestra la nueva distribución tras eliminar las palabras que aparecen en menos del 0.1 % de los documentos de su época. En este nivel de procesamiento se redujo bastante el tamaño del vocabulario a 3,148 (al rededor de 14 veces) sin alterar tan significativamente la cantidad de *tokens* (alrededor de un 10 %), siendo ahora 925,693 *tokens*.

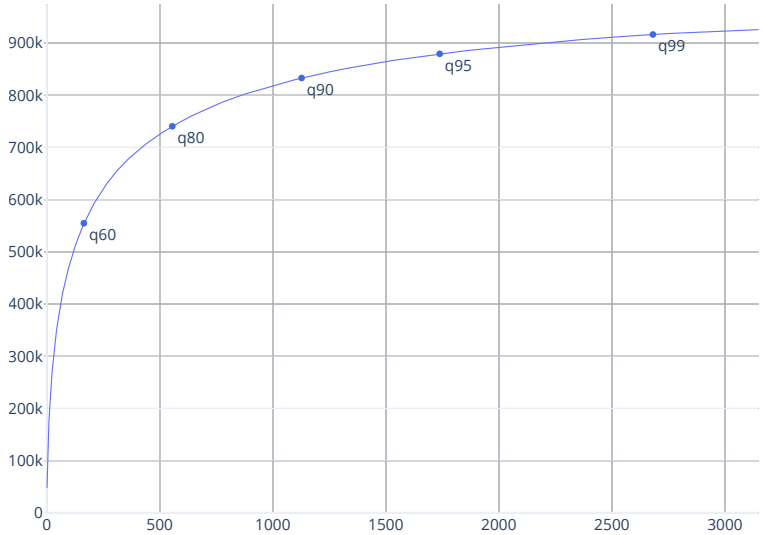


Figura 4.5: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el tercer nivel de procesamiento.

Luego se filtran palabras usando el vocabulario extraído del SUC. En la Figura 4.6 se observa que el vocabulario se redujo a 2,902 y el la cantidad de *tokens* a 901,745. En este

caso la variación no fue tan significativa, alrededor de un 8 % en el tamaño del vocabulario y de un 3 % en el caso de los *tokens*.

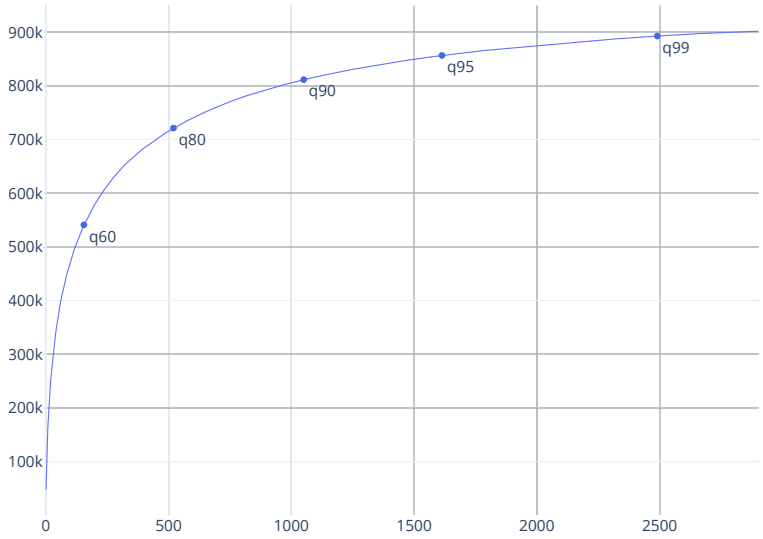


Figura 4.6: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el cuarto nivel de procesamiento.

A continuación, se eliminan las *stopwords*, de la Figura 4.7 se puede observar que esto significó una reducción significativa de tanto el vocabulario como en el número de *tokens*, respectivamente en 32 %(1,960 palabras) y 45 %(495,182 *tokens*). La reducción abrupta en la cantidad de *tokens* se debe principalmente a que las *stopwords* son parte de las palabras más frecuentes dentro del corpus.

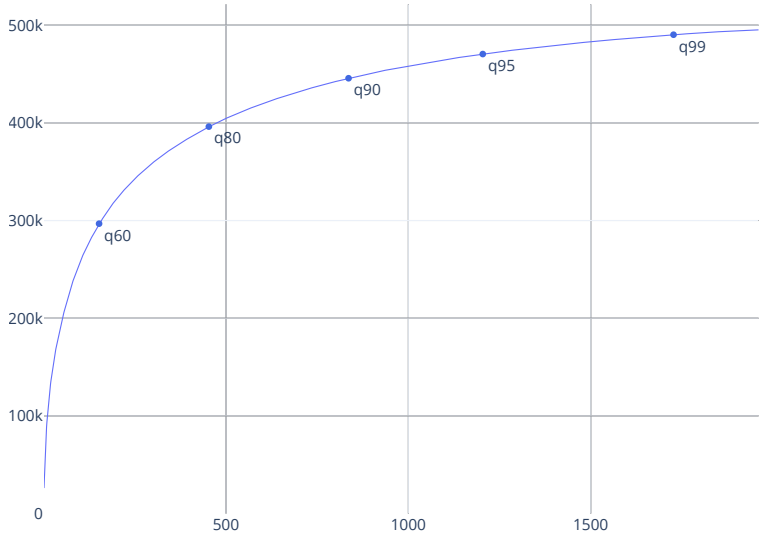


Figura 4.7: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el quinto nivel de procesamiento.

Por último, se eliminan los documentos con menos de 5 palabras, de la Figura 4.8 se puede

observar que esto implicó una reducción de alrededor del 20 % en el tamaño del corpus y del 8 % en el número de *tokens*.

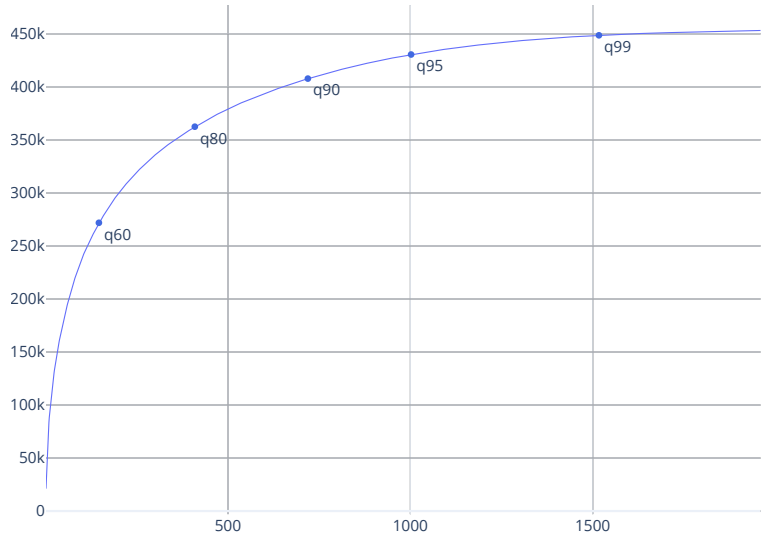


Figura 4.8: Frecuencia acumulada del vocabulario en orden decreciente de ocurrencia aplicando hasta el sexto nivel de procesamiento.

En la tabla 4.1 se muestra un cuadro resumen con estadísticas del corpus bajo distintos niveles de procesamiento. De aquí se extrae que el tamaño del vocabulario, el corpus y la cantidad de *tokens* se redujo en alrededor de un 98 %, un 20 % y un 76 % respectivamente.

procesamiento	documentos	vocabulario	tokens
t	49,015	93,203	2,030,980
t+ch	49,003	42,921	1,028,412
t+ch+f	48,988	3,148	925,693
t+ch+f+v	48,988	2,902	901,745
t+ch+f+v+s	48,566	1,960	495,182
t+ch+f+v+s+d	38,850	1,960	453,206

Tabla 4.1: Estadísticas del corpus bajo distintos niveles de procesamientos, **t**: tokenización, **ch**: procesamiento de caracteres, **f**: filtro por frecuencia, **v**: filtro por vocabulario, **s**: eliminación de *stopwords*, **d**: eliminación de documentos.

En la tabla 4.2 se muestra el detalle del vocabulario para cada una de las épocas tras procesar el corpus, de aquí se extrae que en promedio un 12.83 % del vocabulario se olvida de una época a otra y un 18.92 % es nuevo, es otras palabras, en promedio alrededor de un 32 % del vocabulario no es común entre tópicos de épocas adyacentes. Esto justifica la necesidad por utilizar medidas de similitud robustas a cambios en el vocabulario, permitiendo así una comparación más justa entre tópicos que no tienen un vocabulario común.

época	t-1	t	t-1 [%]	t [%]
2	1,145	1,187	14.41	18.08
3	1,187	1,281	13.56	21.48
4	1,281	1,329	13.35	17.10
5	1,329	1,405	12.57	18.28
6	1,405	1,537	10.25	19.64

Tabla 4.2: Evolución del vocabulario en el tiempo, **t-1**: corresponde al vocabulario del período anterior a la época respectivame,**t-1**: corresponde al vocabulario de la época actual, **t-1**[%] : porcentaje de palabras del período $t-1$ que ya no están en el período t y **t**[%]: porcentaje de palabras del período t que no están en el período $t-1$.

4.3. Análisis cuantitativo de resultados

En esta sección se describen los resultados cuantitativos de aplicar la metodología propuesta al corpus descrito en la sección 4.1. En primer lugar, se analiza el comportamiento temporal de los tópicos bajo distintos puntos operantes ζ para podar el grafo completo. En segundo lugar, se analizan los resultados de aplicar la heurística descrita en 3.3.2, la cual mejora los tiempos de construcción del grafo de similitud a costa de pérdida en precisión.

4.3.1. Grafo de similitud temporal

Como se menciona en la sección 3.3, el grafo temporal se construye relacionando los tópicos de épocas adyacentes a través de una medida de similitud, los tópicos son descubiertos usando el modelo HDP y WMD por medida de similitud. En base a un punto operante $\zeta \in [0, 1]$ de la cdf de la similitud del grafo completo se determina el úmbral de corte. Aquellos arcos con similitud menor a este úmbral son eliminados. De esta forma la elección del úmbral no depende fuertemente de la medida de similitud escogida. En la Figura 4.9 se ilustra la cdf del grafo *fully-connected* obtenido de aplicar la metodología propuesta al corpus descrito en la sección 4.1.

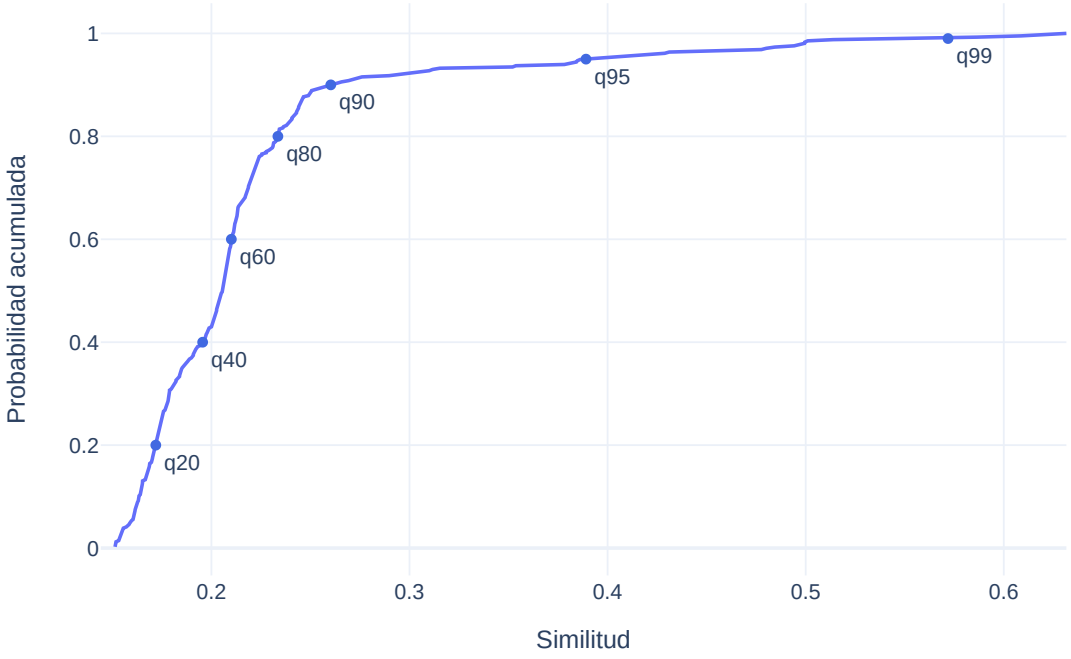
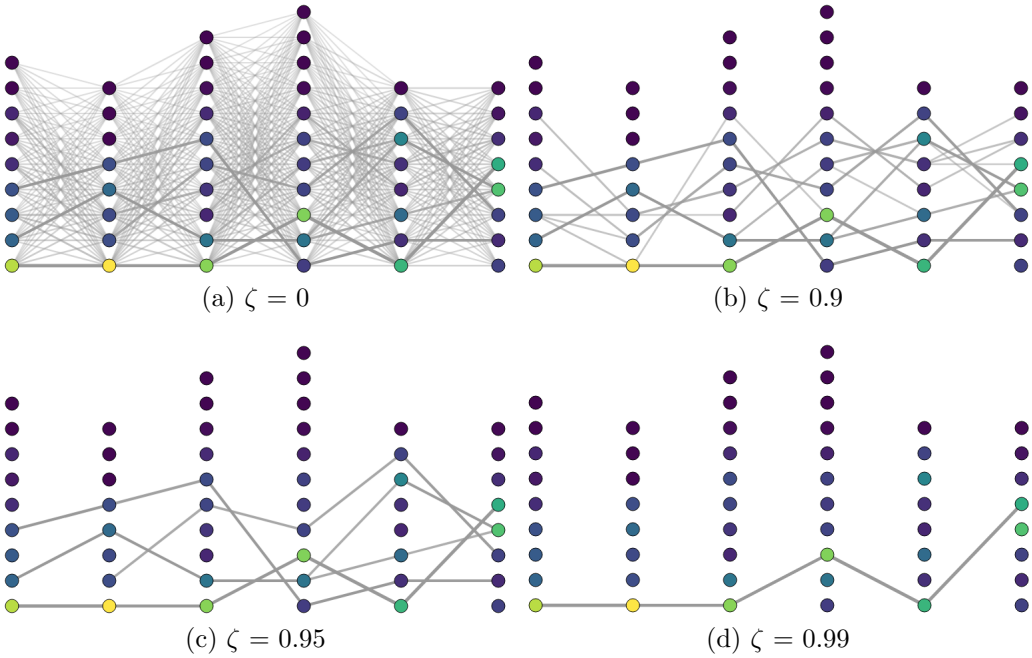


Figura 4.9: Estimación empírica de la cdf de la similitud WMD entre tópicos del grafo *fully conected*.

En la Figura 4.10 se muestran los grafos resultantes tras aplicar distintos puntos operantes ζ. Como es esperado, un incremento de ζ resulta en un incremento en la *sparsity* del grafo temporal final.



(a) $\zeta = 0$ (b) $\zeta = 0.9$

(c) $\zeta = 0.95$ (d) $\zeta = 0.99$

Figura 4.10: Grafo de similitud temporal. Los tres grafos corresponden al mismo corpus y fueron construidos usando el mismo conjunto de tópicos con WMD como medida de similitud, pero bajo diferentes puntos operantes ζ de la CDF. El eje horizontal denota el tiempo en años, partiendo en el 2011 hasta el 2016, donde cada columna de tópicos corresponde a una época específica. Mientras más claro sea el color del nodo que representa un tópico más popularidad posee en su correspondiente época y mientras mayor es el grosor del arco entre dos tópicos mayor es su similitud.

En el grafo podado se pueden identificar diferentes dinamismos en cada época, como nacimiente, muerte, evolución, fusión y división. Un tópico nace en una época si en la época anterior no posee antecesores. La época de muerte de un tópico se identifica como aquella en que no tiene ningún descendiente. Un tópico evoluciona si posee un único antecesor. La fusión ocurre en una época cuando un tópico tiene más de un antecesor. En cambio, la división ocurre cuando hay dos o más tópicos de una época que comparten un antecesor.

En la Figura 4.11 se muestra la proporción de tópicos que nacen, mueren, fusionan, y dividen por época, normalizando por el número total de tópicos inferido en esa época. Se puede observar que al incrementar ζ la proporción de tópicos que nacen y mueren aumenta significativamente, puesto que al ser mayor el úmbral de corte menos ascendentes y descendientes se tienen, por otro lado, la división y fusión decrecen, debido a la menor presencia de conexiones.

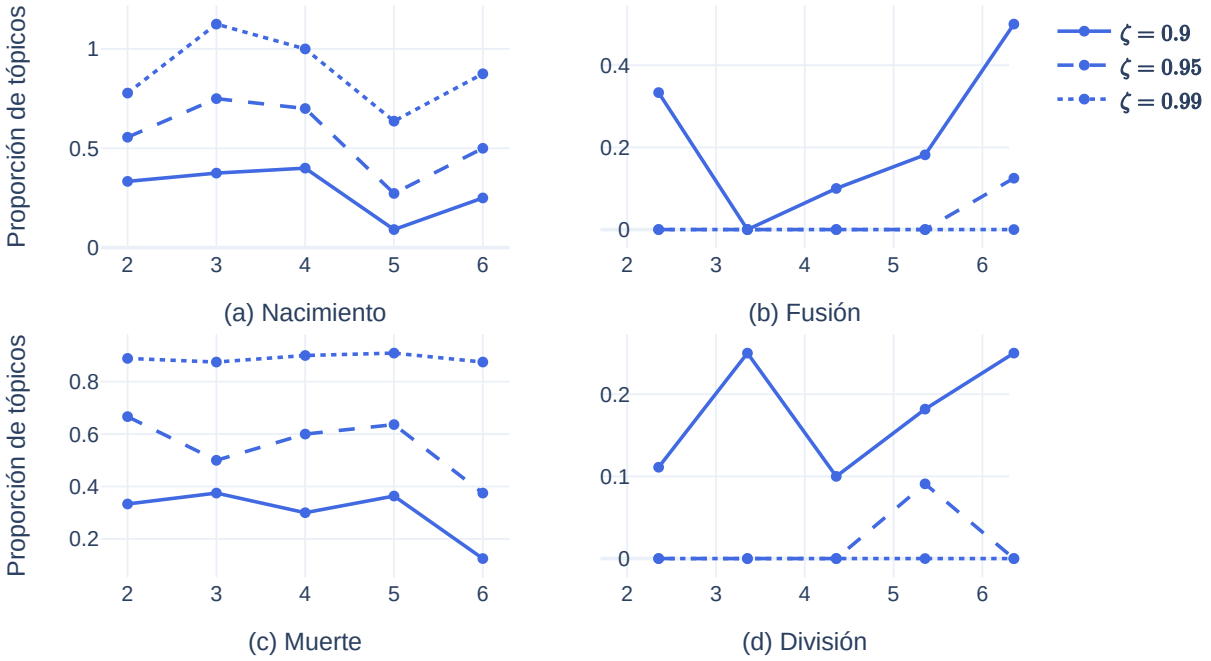


Figura 4.11: Proporción de tópicos que nacen, mueren, fusionan y dividen por época, normalizando por el número total de tópicos inferido en esa época, bajo diferentes puntos operantes ζ.

4.3.2. Heurística de mejora del tiempo de construcción del grafo de similitud

WMD es una medida intensiva en recursos computacionales, por lo que se requiere de herísticas para escalar la metodología a un gran volumen de datos. Como se menciono en la sección 3.3.2 una forma de reducir significativamente el tiempo de creación del grafo temporal es utilzar el top N de palabras más probables del tópico que expliquen determinado porcentaje de la distribución acumulada. A la hora de aplicar esta heurística se debe tener en cuenta la posible pérdida de información al representar un tópico con un vocabulario reducido. Sea G_ζ el grafo obtenido al aplicar tras podar el grafo completo con un punto operante ζ y G'_ζ el grafo aproximado de aplicar la heurística, con E y E' los conjuntos de arcos respectivos. Luego, el porcentaje de arcos correctos de la heurística corresponde a la cardinalidad de la intersección divida por la cardinalidad de la unión, matemáticamente:

$$\frac{|E \cap E'|}{|E \cup E'|} \tag{4.1}$$

En la Figura 4.12 se muestra la mejora en *speedup* y el incremento en error al usar un porcentaje menor de la cdf del tópico para construir el grafo de similitud temporal. Utilizando

el 20 % de la cdf la construcción del grafo de similitud se acelera en más de 1,207 veces, esto bajo un error medianamente aceptable del orden del 10-30 %. Si se utiliza un 60 % de la cdf se obtiene una aceleración casi 10 veces menor, específicamente de 137, siendo el error menor al 10 %. Por otro lado, si se utiliza un 90 % de la cdf el error es inferior al 5 %, pero bajo un aceleración de 8 superior a utilizar el vocabulario completo del tópico en la construcción del grafo temporal.

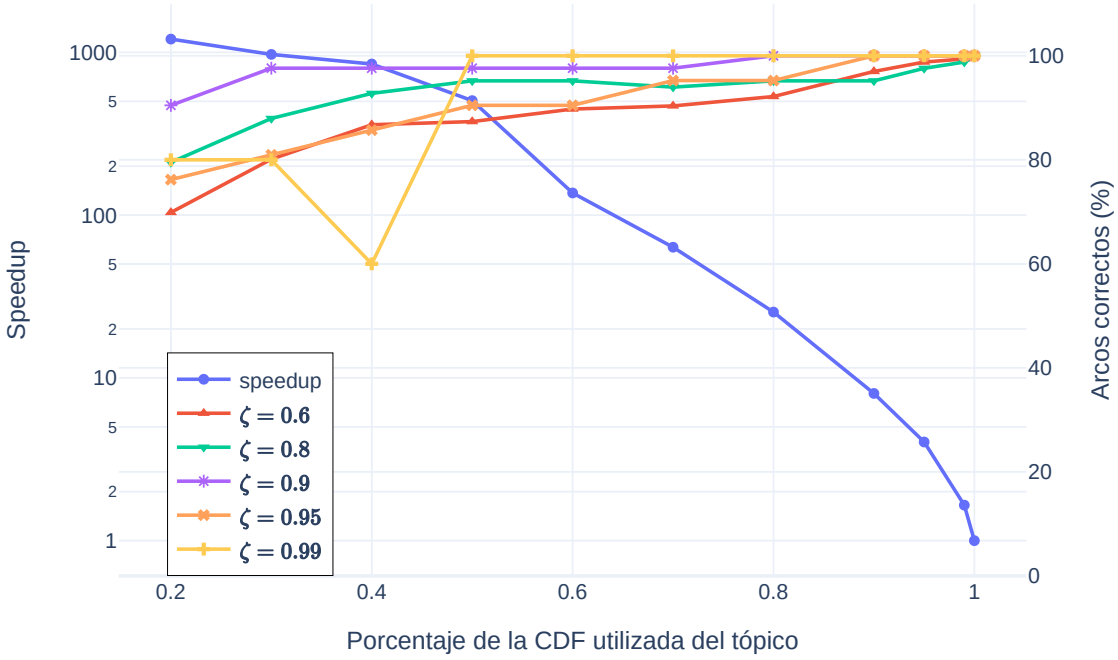


Figura 4.12: *Speedup* y porcentaje de arcos correctos al utilizar un menor porcentaje de la cdf de los tópicos en la construcción del grafo de similitud. El error de la heurística es mostrado para diferentes puntos operantes ζ utilizados para podar el grafo completo.

4.4. Análisis cualitativo de resultados

En esta sección se realiza un análisis cualitativo de los tópicos descubiertos y su evolución. El análisis es realizado sobre el grafo podado con un punto operante $\zeta = 0.95$ mostrado en la Figura 4.10. En las siguientes dos subsecciones se analizan los dos tópicos más predominantes en el tiempo: en la sección 4.4.1 se estudia el robo no presencial y en la sección 4.4.2 se analiza el robo con violencia.

4.4.1. Evolución del robo no presencial

En la Figura 4.13 se puede muestra la evolución de uno de los tópicos más predominantes del grafo temporal. A diferencia de otros tópicos este tópico no presenta en el tiempo división, fusión ni desaparición. En base a las top 10 palabras más probables el tópico se puede caracterizar como un robo bajo las siguientes circunstancias: vehículo estacionado en la calle, sin la presencia del conductor, el conductor al regresar al lugar se percata de la ausencia del vehículo. Un nombre adecuado para este tópico podría ser "robo no presencial", ya que se caracteriza por no contar con la presencia del conductor cuando ocurre el robo. Este tópico se caracteriza por su gran estabilidad en el tiempo, ya que en el top 10 se encuentran practicamente las mismas palabras. Por último, se observa una tendencia a la baja de la participación del tópico, en el año 2011 alrededor del 44 % de los *tokens* provienen de este tópico y en el año 2016 se tiene un 32 %.

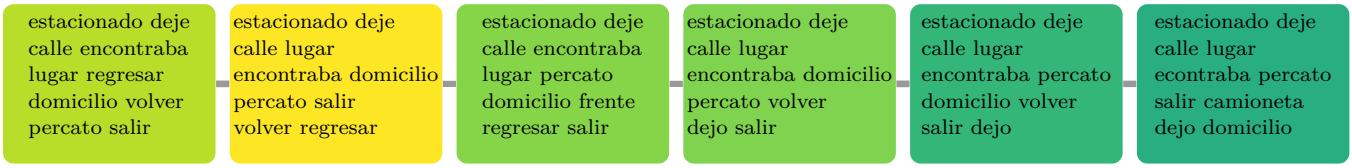


Figura 4.13: Evolución del tópico de robo de vehículo no presecial. El eje horizontal denota el tiempo en años, partiendo en el 2011 hasta el 2016. Mientras más claro sea el color del tópico más popularidad posee en su correspondiente época y mientras mayor es el grosor del arco entre dos tópicos mayor es su similitud.

4.4.2. Evolución del robo con violenca

En la Figura 4.14 se describe un tipo de robo de vehículo que se podria clasificar como robo con violencia. Este delito, se presentaría de la siguiente forma: un conjunto de personas se bajan de un vehículo, intimidan al conductor mediante una pistola u otra arma de fuego, le quitan las llaves del vehículo y finalmente se llevan el vehículo dándose a la fuga. La participación de este tipo de robo se ha visto al alza, en el 2011 su participación era del 12 % y en el 2016 del 36 %, por lo que se ha convertido en un delito más a la moda, quitandole terreno al robo no presencial. A diferencia del tópico robo no presencial este presenta una fusión y división en el 2015. En ese año emerge del robo con violencia un "subtópico" popularmente conocido como "portonazo", un robo con violencia que se caracteriza por la sustracción del vehículo en el portón de la casa de la víctima. Ese año coincide con la popularización de aquel modus operandi. Al siguiente año esste subtópico se vuelve a fusionar para formar el tópico robo con violencia de períodos anteriores.

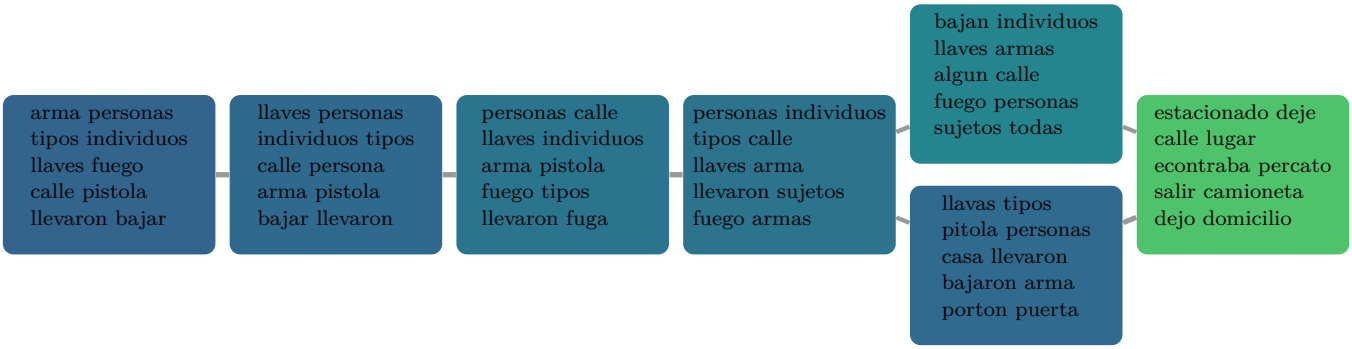


Figura 4.14: Evolución del tópico de robo con violencia de vehículo. El eje horizontal denota el tiempo en años, partiendo en el 2011 hasta el 2016. Mientras más claro sea el color del tópico más popularidad posee en su correspondiente época y mientras mayor es el grosor del arco entre dos tópicos mayor es su similitud.

Capítulo 5

Conclusiones y trabajo futuro

Este trabajo se enfoca en el modelamiento y descubrimiento de tópicos en el tiempo en un corpus. El enfoque descrito comienza con la discretización del corpus en épocas. Luego, usando como aproximación que la estructura de los tópicos dentro de cada época es estática, los tópicos son descubiertos usando Hierarchical Dirichlet Process. Por último, la evolución de los tópicos en el tiempo es modelada por un grafo temporal sustentado por una medida de similitud entre tópicos. El grafo inicialmente es construido por los arcos entre todos los pares de tópicos de épocas adyacentes, luego es podado automáticamente en base a un punto operante de la cdf de la similitud. Esta estructura permite inferir cambios estructurales complejos de los tópics, como nacimiento, muerte, división, fusión y evolución en el tiempo.

En contraste a trabajos anteriores, la metodología propuesta utiliza Word Mover's Distance como medida de similitud entre tópicas, permitiendo comparar de forma más apropiada tópicos que no poseen un vocabulario común a través de sus *word embeddings*. Además, se presenta una análisis empírico del trade off entre precisión y *speedup* de no utilizar el vocabulario completo del tópico en la construcción del grafo de similitud.

Resultados experimentales al fenómeno de robo de vehículos son reportados. Se muestra el efecto que tiene la elección del punto operante de la cdf en la construcción del grafo final. Otro importante hallazgo, es nivel de consistencia de la metodología en modelar de la evolución de los tópicos en el tiempo, siendo capaz de relacionar tópicos que evidentemente son similares, como es el caso del robo no presencial o el robo con violencia.

La metodología propuesta puede tener otros usos interesantes como descubrir nuevas tendencias de investigación, analizar la evolución de la contigencia social, estudiar la efectividad de campañas publicitarias en base a la opinión de los consumidores, organizar y recomendar contenido en un blog, etc.

Como trabajo futuro podría ser interesante extender la metodología a un enfoque puramente basado en redes neuronales. De esta forma la comparación entre tópicos de épocas adyacentes a través de sus *word embeddings* se vuelve más natural. También sería más consistente ya que la información codifica en los *word embeddings* se utilizaría para el mismo descubrimiento de los tópicos. En Dieng et al. (2019) se propone una modelo de tópicos dinámico basado en redes neuronales, sin embargo, este mantiene fijo el número de tópicos en el tiempo.

Bibliografía

Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

Amr Ahmed and Eric P Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

Andrew T Wilson and David G Robinson. Tracking topic birth and death in lda. *Sandia National Laboratories*, 2011.

Adham Beykikhoshk, Ognjen Arandjelović, Dinh Phung, and Svetha Venkatesh. Discovering topic structures of a temporally evolving document corpus. *Knowledge and Information Systems*, 55(3):599–632, 2018.

Thomas Minka. Estimating a dirichlet distribution, 2000.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Erik Blaine Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, 2006.

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3): 103–134, 2000.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

W John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of information science*, 18(1):45–55, 1992.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. .ºReilly Media, Inc.", 2009.

Chong Wang and David Blei. HDP: Hierarchical dirichlet process C++, 2010. URL https://github.com/blei-lab/hdp.

Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Vlad Niculae. Word mover's distance in python, 2015. URL http://vene.ro/blog/word-movers-distance-in-python.html.

Gary Doran. PyEMD: Earth mover's distance for Python, 2014. URL https://github.com/garydoranjr/pyemd.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

José Cañete. Fasttext embeddings from SUC. https://github.com/BotCenter/spanishWordEmbeddings, 2019a.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

José Cañete. Spanish Unannotated Corpora, 2019b. URL https://github.com/josecannete/spanish-corpora.

Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

Romain Deveaud, Eric SanJuan, and Patrice Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014.

Wen Zhang, Yangbo Cui, and Taketoshi Yoshida. En-lda: An novel approach to automatic bug report assignment with entropy optimized latent dirichlet allocation. *Entropy*, 19(5): 173, 2017.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019.