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 Ingeniería Civil Industrial

Diego Garrido

Profesor guía: Richard Weber
Miembros de la comisión: Ángel Jiménez, Giorgiogiulio Parra

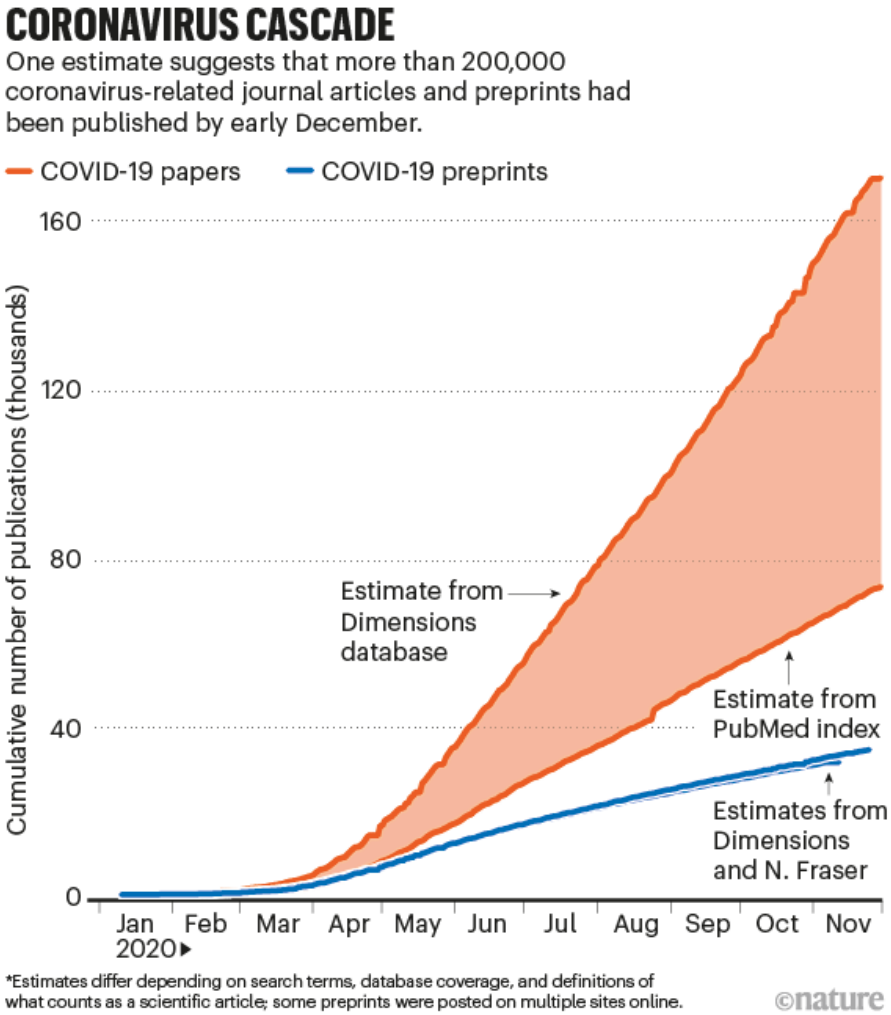
Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Industrial

12 de enero de 2022

Contenidos

Motivación: Metodología

- Incremento de volumenes de datos requiere métodos automáticos de análisis.
- El modelamiento de tópicos permite resumir y organizar grandes colecciones de texto.
- Este trabajo es una propuesta de modelamiento de tópicos dinámico: nacimiento, muerte, evolución, división y fusión de tópicos.
- En contraste a trabajos anteriores, la metodología incorpora WMD como medida de similitud para comparar tópicos sin vocabulario común.



Motivación: Caso de estudio

El problema del robo de vehículos o accesorios de vehículos afecta a toda la sociedad en Chile y en el mundo.

Algunos de sus efectos son:

- Incremento en la percepción de seguridad.
- Aumento en la prima de los seguros de los asegurados.
- Aumento en los costos de las aseguradoras.
- Incremento en otro tipos de delitos.

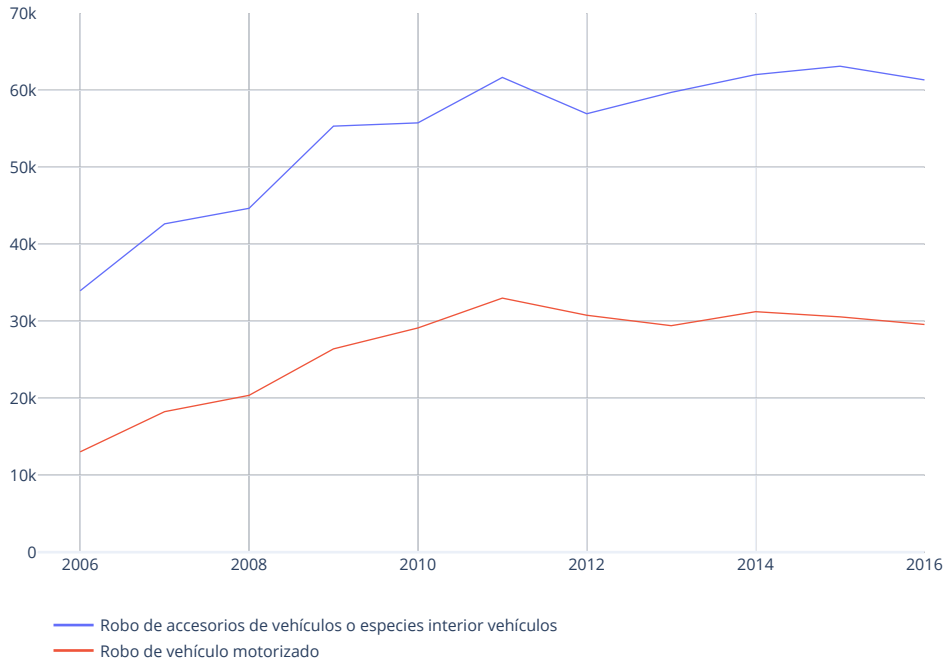


Figura 1: Cantidad de robos de vehículos y accesorios anuales en Chile entre los años 2006-2016. Fuente: Informe anual de Carabineros, 2006-2016, INE.

Revisión del estado del arte: ¿Qué es el modelamiento de tópicos?

El modelamiento de tópicos es uno de los enfoques más prometedores 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.

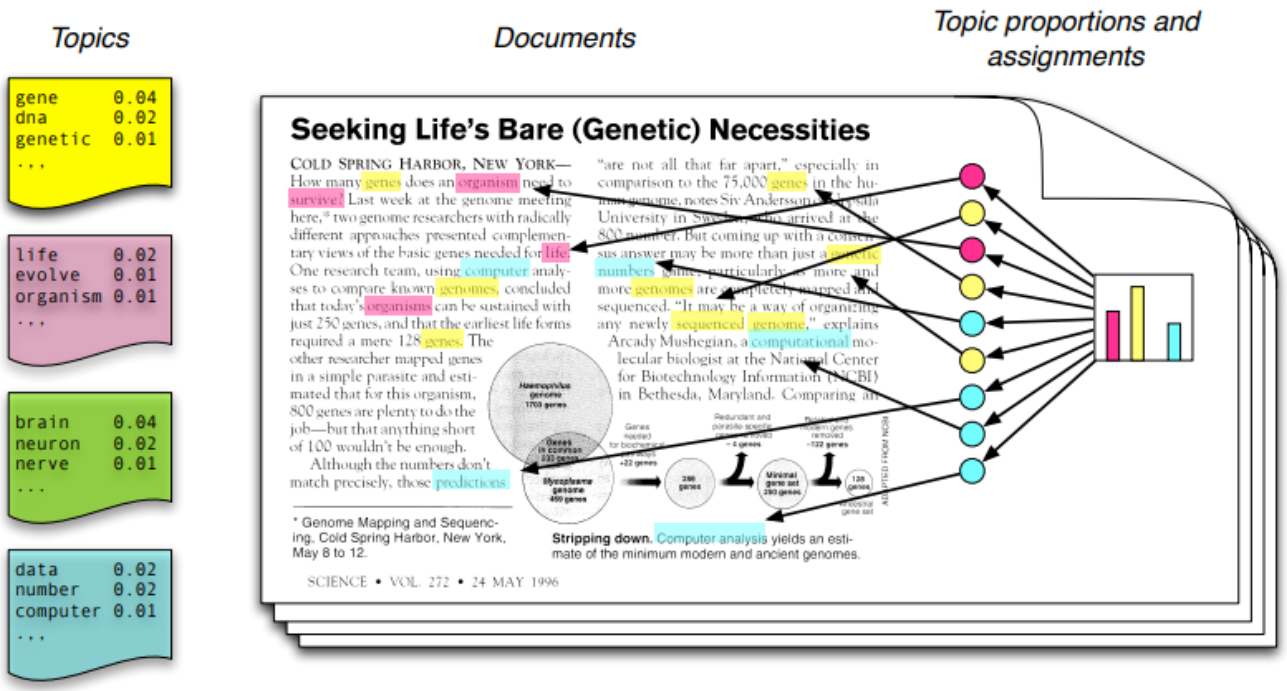


Figura 2: Intuición detrás de Latent Dirichlet Allocation. Fuente: Figura 1 de [Blei, 2012].

Revisión del estado del arte: Tipos de modelos de tópicos

Las técnicas de modelamiento de tópicos suelen estar basadas en **factorización matricial** o en **modelos probabilísticos generativos**.

A continuación algunos ejemplos de ambos enfoques:

- **LSI** (Latent Semantic Indexing) [Dumais, 2004] o **NMF** (Non-negative Matrix Factorization) [Xu et al., 2003].
- **LDA** (Latent Dirichlet Allocation) [Blei et al., 2003] o **HDP** (Hierarchical Dirichlet Process) [Teh et al., 2005].

Este trabajo aborda el enfoque probabilístico:

- **Expresa incertidumbre** en la asignación de un tópico a un documento y en la asignación de palabras a los tópicos.
- Suele aprender **tópicos más descriptivos** [Stevens et al., 2012].

Revisión del estado del arte: Modelamiento dinámico

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

1. **Evolución de tópicos.**
2. **Dinámismo en la mezcla de tópicos.**
3. **Nacimiento, muerte, fusión y división de tópicos.**

Dentro de los modelos de tópicos dinámicos se tiene:

- **DTM** (Dynamic Topic Modelling) [Blei and Lafferty, 2006] y **TOC** (Topic Over Time) [Wang and McCallum, 2006] **permiten el punto 1 y 2** manteniendo fijo el número de topicos en el tiempo.
- **DHDP** (Dynamic Hierarchical Dirichlet Process) [Ahmed and Xing, 2012] **captura el punto 1, 2 y 3 parcialmente**, con excepción de fusión y división. **No cuenta con una implementación**.
- En [Wilson and Robinson, 2011] y [Beykikhoshk et al., 2018] se capturan los dinámismos mencionados dividiendo el corpus en épocas, entrenando de forma independiente un modelo por época para finalmente unificar (LDA y HDP).

Metodología propuesta

1. División del corpus en épocas siendo cada época preprocesada de forma independiente.
2. Aplicación de HDP en cada época de manera independiente.
3. Construcción del grafo de similitud WMD entre épocas adyacentes y eliminación de arcos con similitud menor al cuantil ζ de la **cdf**.

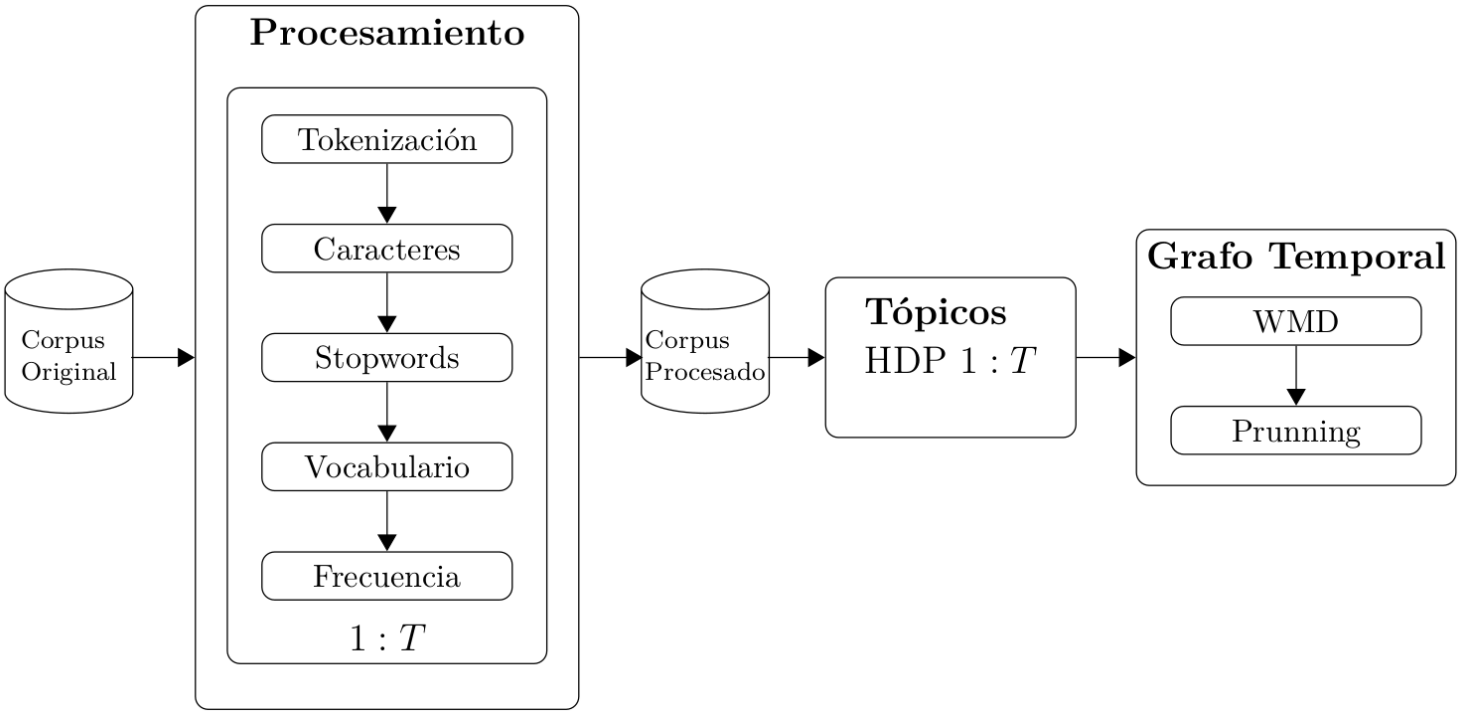


Figura 3: Esquema de la metodología de descubrimiento y evolución de tópicos.

Metodología propuesta: Hierarchical Dirichlet Process

HDP (Hierarchical Dirichlet Process) es un *prior* jerárquico no paramétrico formado por un DP cuya medida base G_0 es dibujada a partir de otro DP.

$$
\begin{aligned}
H &= \text{Dir}\left(\frac{\eta}{|V|}1_{|V|}\right) \\
G_0|\gamma, H &\sim \text{DP}(\gamma, H) \\
G_d|\alpha_0, G_0 &\sim \text{DP}(\alpha_0, G_0) \\
\phi_{d,n}|G_d &\sim G_d \\
w_{d,n}|\phi_{d,n} &\sim \text{Cat}(\phi_{d,n})
\end{aligned}
$$

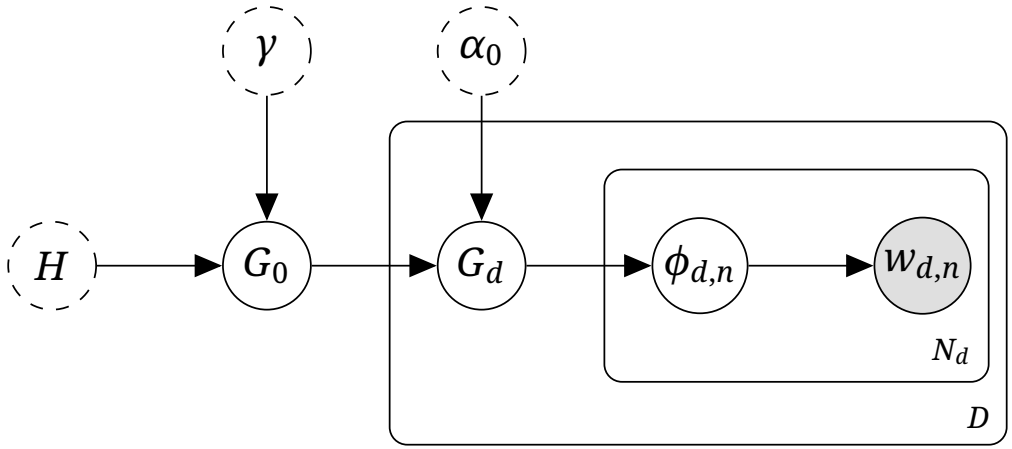


Figura 4: Representación gráfica de HDP: círculos denotan variables aleatorias, círculos abiertos denotan parámetros, círculos sombreados denotan variables observadas y los platos indican replicación.

La discretitud de G_0 asegura:

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

Metodología propuesta: Construcción del grafo de similitud temporal

- Construcción del grafo *fully connected* de las similitudes entre tópicos de épocas adyacentes ($\phi_{t,i}$ y $\phi_{t+1,j}$) usando una medida de similitud $\rho \in [0, 1]$.
- Eliminación de las conexiones débiles en base a un umbral $\alpha \in [0, 1]$, reteniendo solo aquellas conexiones que cumplen $\rho(\phi_{t,i}, \phi_{t+1,j}) > \alpha$.

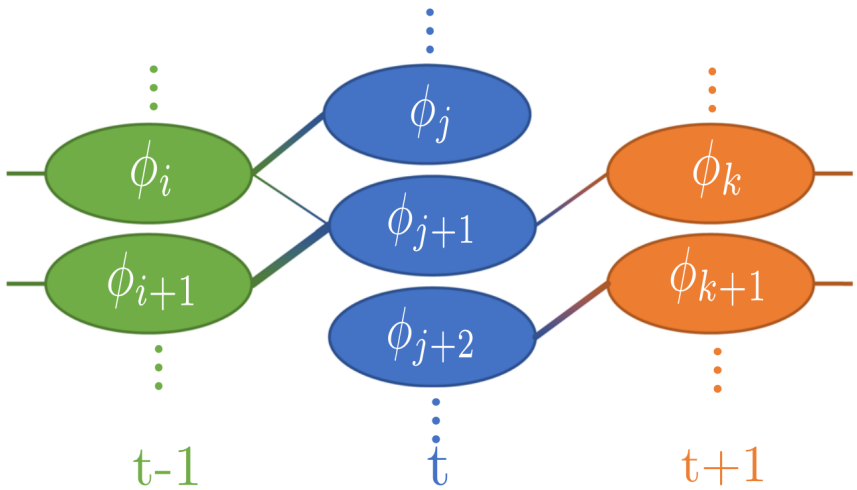


Figura 5: Ilustración conceptual del grafo de similitud que modela la dinámica de los tópicos en el tiempo. Un nodo corresponde a un tópico en una época específica; el ancho de los arcos es proporcional a la similitud entre los tópicos, arcos ausentes fueron eliminados por presentar una similitud menor a un umbral. Fuente: Figura 3 de [Beykikhoshk et al., 2018].

Metodología propuesta: Word Mover's Distance

- **WMD** (Word Mover's Distance) [Kusner et al., 2015]: distancia que permite comparar vectores sin vocabulario común ya que trabaja sobre el espacio de los *word embeddings*.
- Sea V_i y V_j los vocabularios del tópico i y j respectivamente, luego se tiene $WMD(\phi_i, \phi_j)$:

$$\min_{x} \sum_{u \in V_i} \sum_{v \in V_j} c_{u,v} x_{u,v} \quad (1)$$

$$\text{s.t.} \sum_{v \in V_j} x_{u,v} = \phi_{i,u}, \ u \in V_i \quad (2)$$

$$\sum_{u \in V_i} x_{u,v} = \phi_{j,v}, \ v \in V_j \quad (3)$$

$$x_{u,v} \geq 0, \ u \in V_i \ , v \in V_j \quad (4)$$

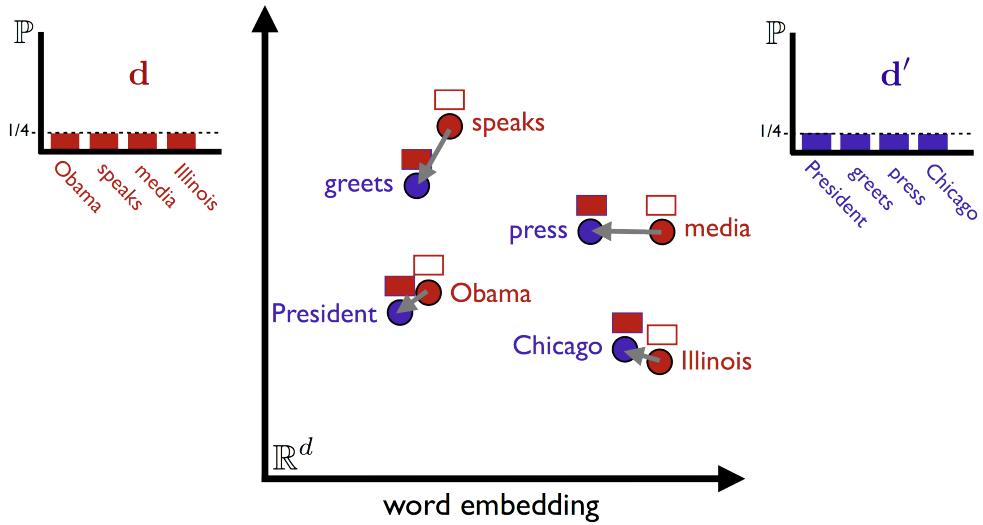


Figura 6: Espacio vectorial de los *word embeddings* de las palabras de dos documentos con un vocabulario de tamaño 4. Fuente: Figura de [Niculae, 2015].

WMD se puede transformar en médida de similitud[1] considerando $\rho(\phi_i, \phi_j) = \frac{1}{1+WMD(\phi_i,\phi_j)}$.

[1] Notar que la similitud es 1 si WMD es 0 y 0 si es ∞.

Metodología propuesta: WMD complejidad

WMD es una medida de distancia **intensiva en recursos computacionales**.

Usando el algoritmo desarrollado por [Pele and Werman, 2009] se tiene que el mejor tiempo promedio escala $O(N^2 logN)$, donde N es el tamaño de la unión los vocabularios de dos épocas adyacentes.

$$\{x|Ax = b, x \geq 0\}, A \in \mathbb{R}^{2N \times N^2}, b \in \mathbb{R}^{2N}, x \in \mathbb{R}^N$$

Se requiere de **heurísticas** para acelerar el tiempo de computo.

- 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.
- En la práctica **la interpretación de los tópicos se basa en los top N palabras más probables**, usualmente con $N \in [5.30]$.

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.

Metodología propuesta: Prunning

- El umbral de corte es el cuantil $\zeta \in [0, 1]$ de la cdf de las similitudes (F_p), es decir, $F_p^{-1}(\zeta)$.
- El úmbral de corte no es arbitrario según la médida de similitud escógida.

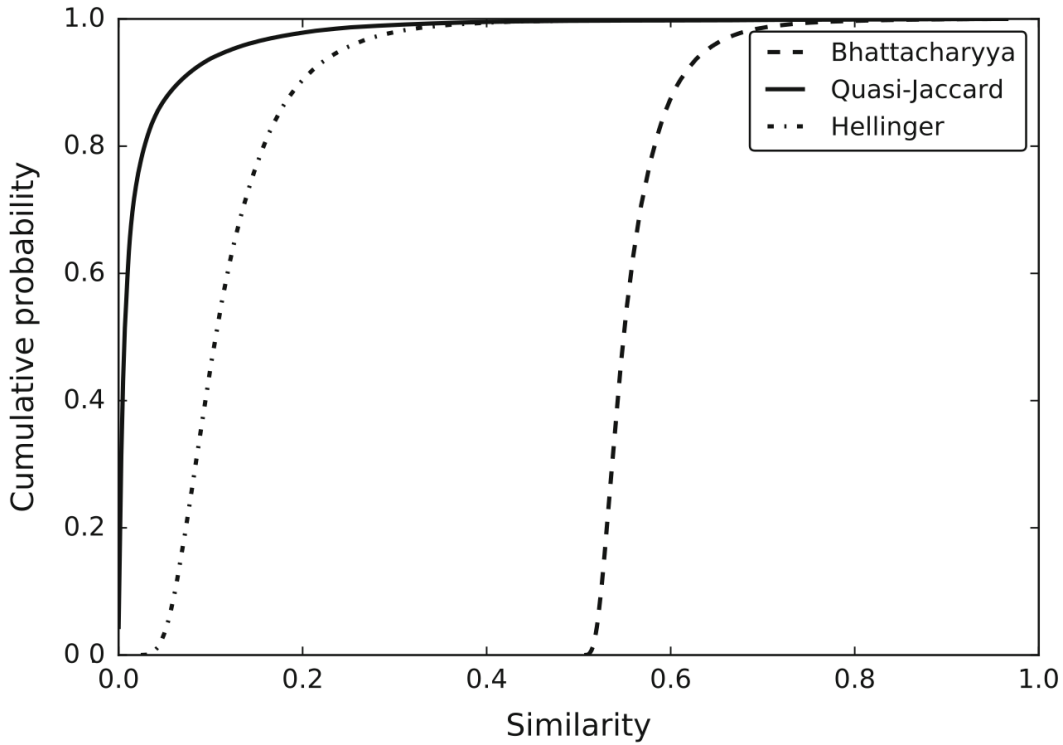


Figura 7: Estimación empírica de la función de densidad acumulada (cdf) de la similitud entre tópicos de épocas adyacentes en un grafo *fully-connected* para tres medidas de similitud. Fuente: Figura 4 [Beykikhoshk et al., 2018].

Metodología propuesta: Hiperparámetros

HDP cuenta con tres hiperparámetros:

- El **parámetro de concentración a nivel corpus** γ y el **parámetro de concentración a nivel documento** α_0. En [Teh et al., 2005] los parámetros de concentración se integran afuera usando un prior *vague gamma* [Escobar and West, 1995]. En este caso se utilizó un prior $\Gamma(\alpha = 1, \beta = 1)$.
- El **parámetro de la medida base Dirichlet** η. Se prefiere usar $\eta \in (0, 1)$ ya que genera distribuciones *sparse* sobre el vocabulario. En este caso se utilizó un punto intermedio, fijando $\eta = 0.5$.

El grafo temporal cuenta con dos hiperparámetros:

- $q \in [0, 1]$ **cuantil de corte de la cdf del tópico**. Se prefieren valores en $[0.8, 0.95]$ ya que conservan el *core* de palabras del tópico y disminuye significativamente el tiempo de cómputo.
- $\zeta \in [0, 1]$ **cuantil de corte de la cdf de las similitudes del grafo fully connected**. Se prefieren valores en $[0.9, 0.99]$ ya que se conservan aquellas relaciones con alta similitud relativa.

Caso de estudio: Corpus

El corpus provisto por la Asociación de Aseguradores de Chile (AACH) consta de 49,015 relatos de víctimas del robo de vehículos provistos entre los años 2011-2016.

> 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

Figura 8: Muestra de relatos de la base de datos AACH.

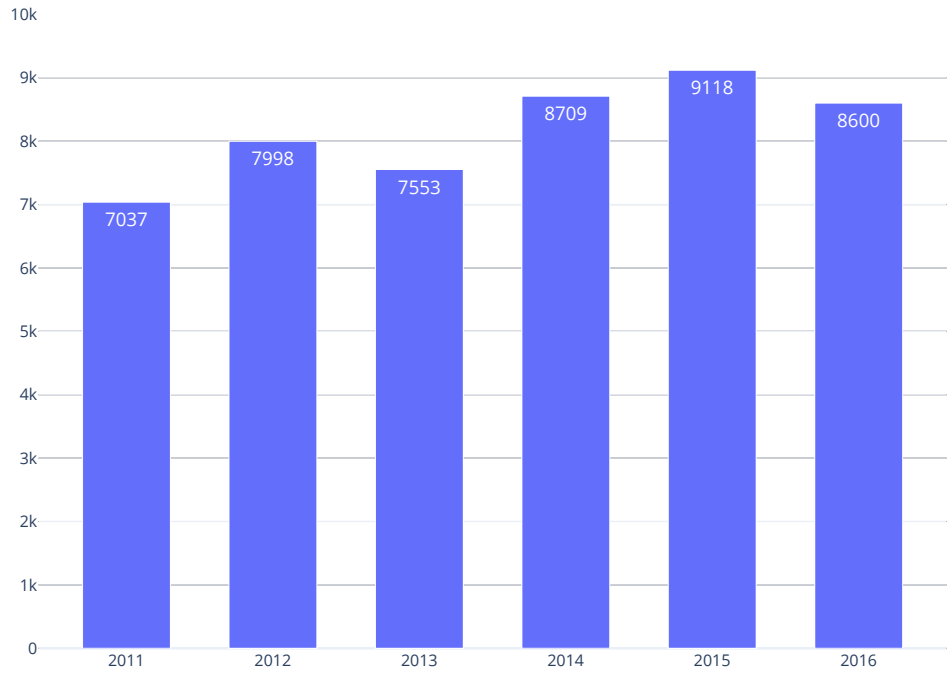


Figura 9: Cantidad de robos registrados por año en base de datos AACH.

Caso de estudio: Procesamiento I

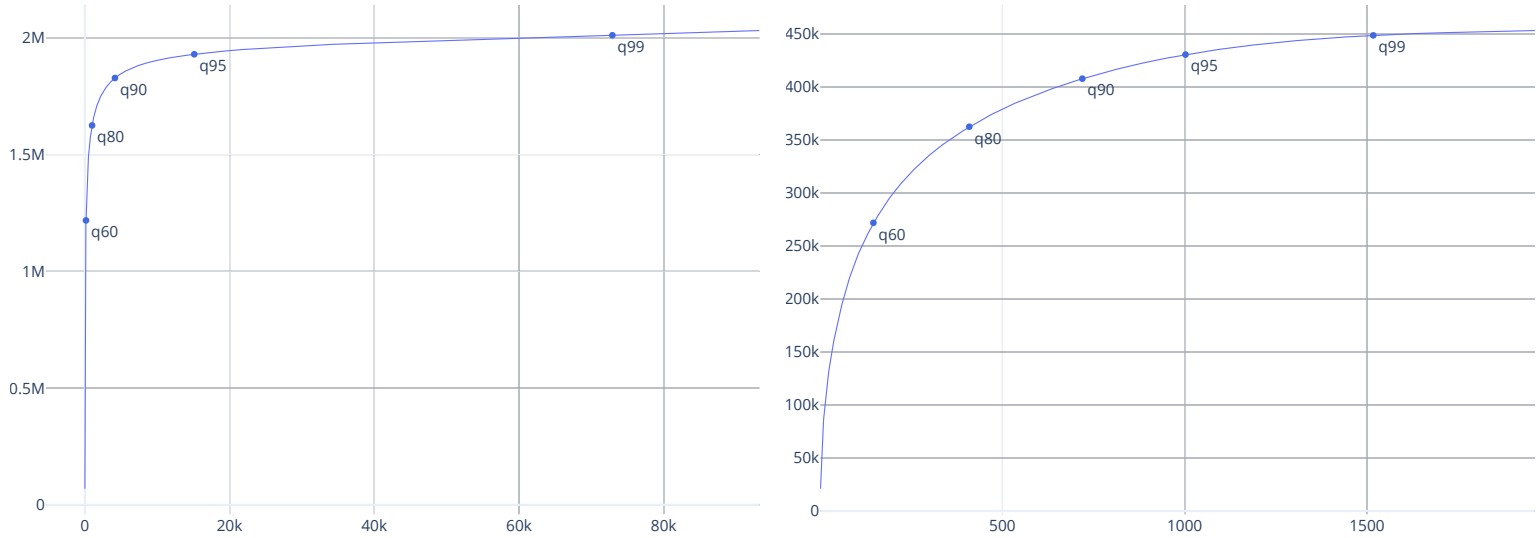
Al procesar el corpus se obtiene una reducción de:

- 20 % del corpus.
- 98 % del vocabulario.
- 76 % de tokens.

Tabla 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.

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

Caso de estudio: Procesamiento II



(a) **(b)**

Figura 10: Distribución acumulada del vocabulario (a) pre-procesamiento y (b) post-procesamiento.

Caso de estudio: Procesamiento III

- En promedio un 18.92 % del vocabulario de una época es nuevo y un 12.83 % se olvida.
- En promedio alrededor de un 32 % del vocabulario no es común entre tópicos de épocas adyacentes.

Tabla 2: Evolución del vocabulario en el tiempo. **t**: vocabulario de la época actual, **t-1**: vocabulario del período anterior, **t[%]**: porcentaje del vocabulario de t que es nuevo y **t-1[%]**: porcentaje del vocabulario de $t-1$ que se olvida.

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

Caso de estudio: CDF del grafo fully-connected

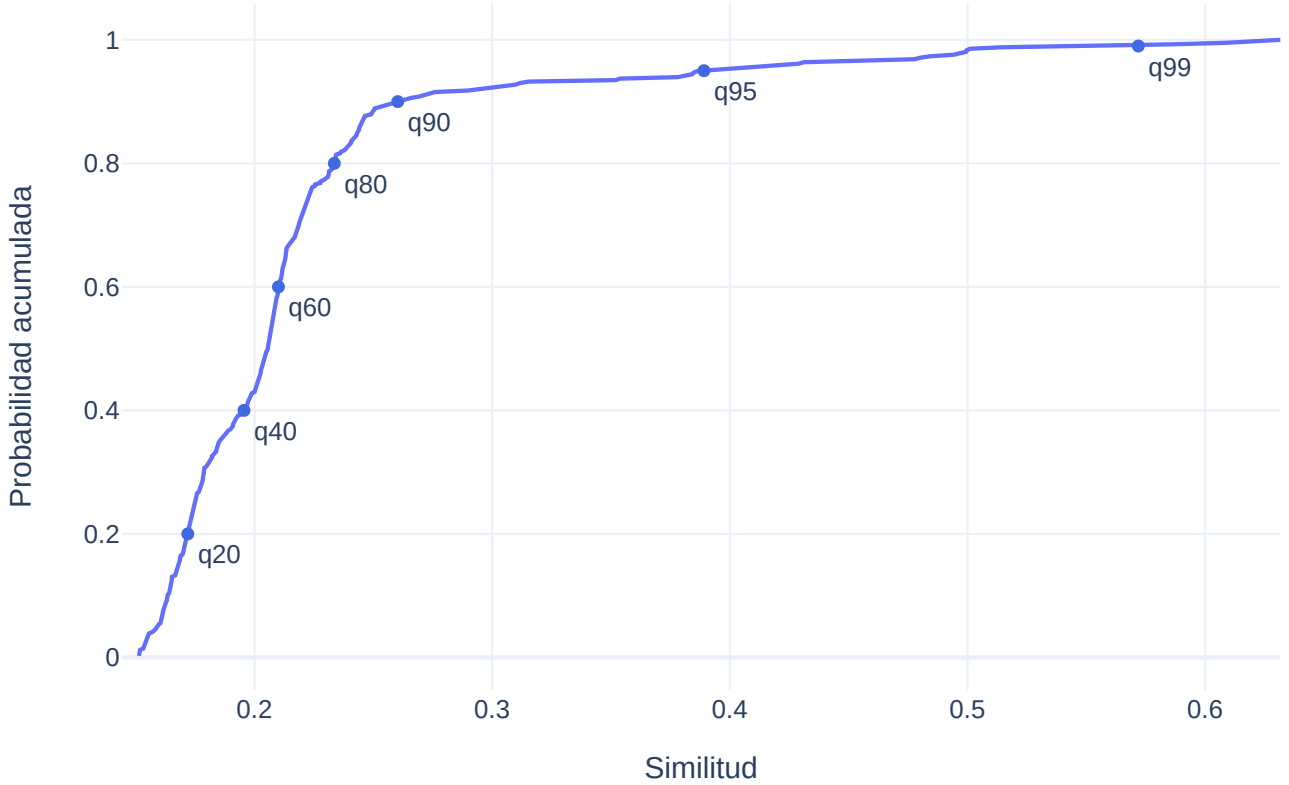
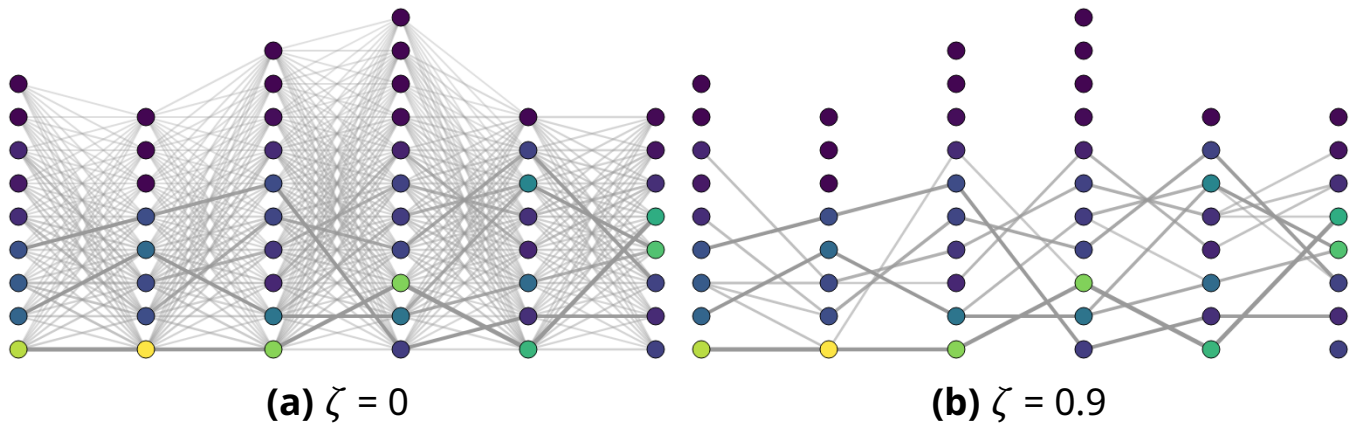
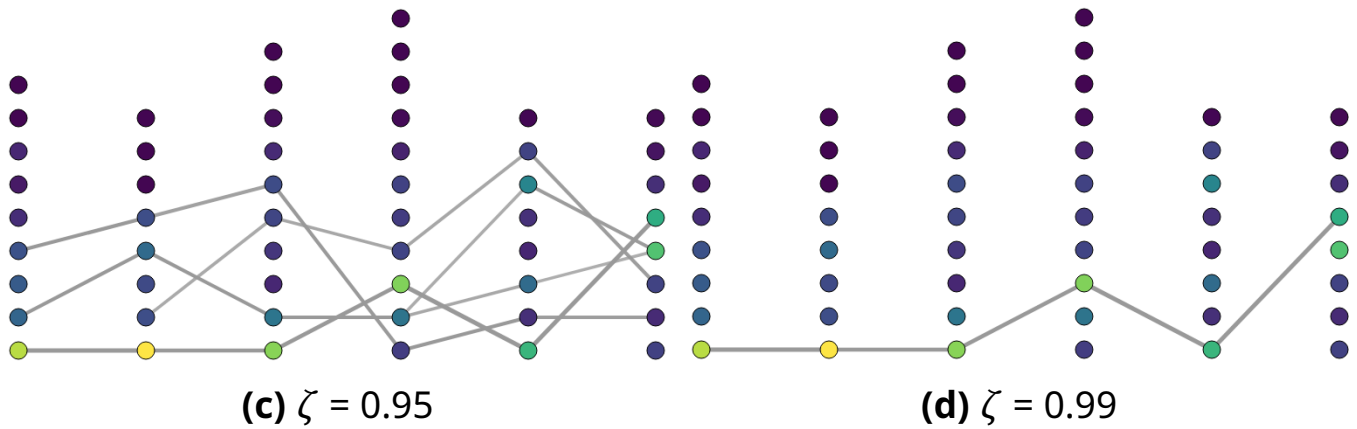


Figura 11: Estimación empírica de la cdf de la similitud WMD entre tópicos del grafo *fully-conected*.

Caso de estudio: Grafo de similitud temporal



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



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

Figura 12: Grafos temporales con WMD como medida de similitud 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.

Caso de estudio: Análisis de sensibilidad del punto operante

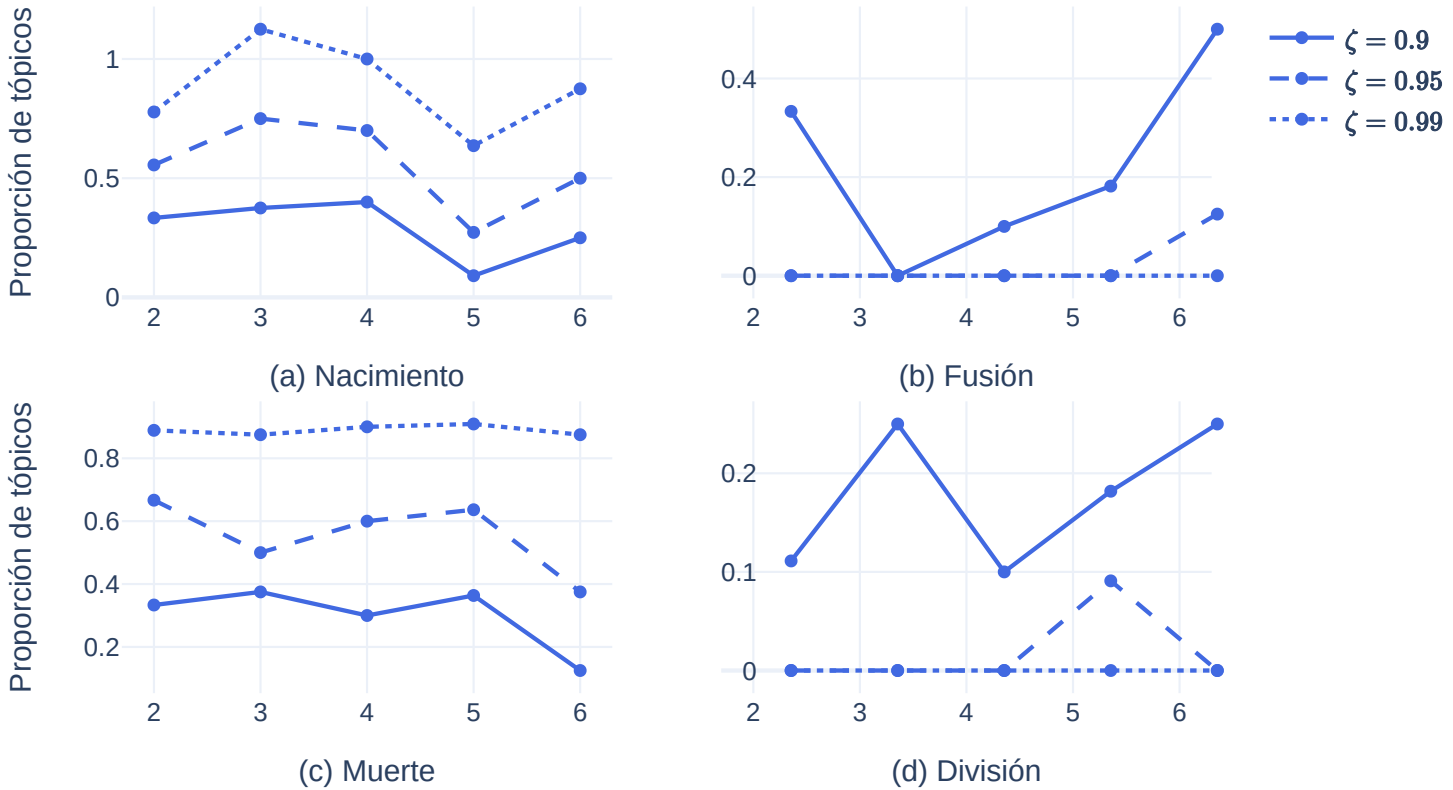


Figura 13: Proporción de tópicos que nacen, mueren, fusionan y dividen por época, normalizado por el número total de tópicos inferido en esa época, bajo diferentes puntos operantes ζ.

Caso de estudio: Trade off entre precisión y speedup

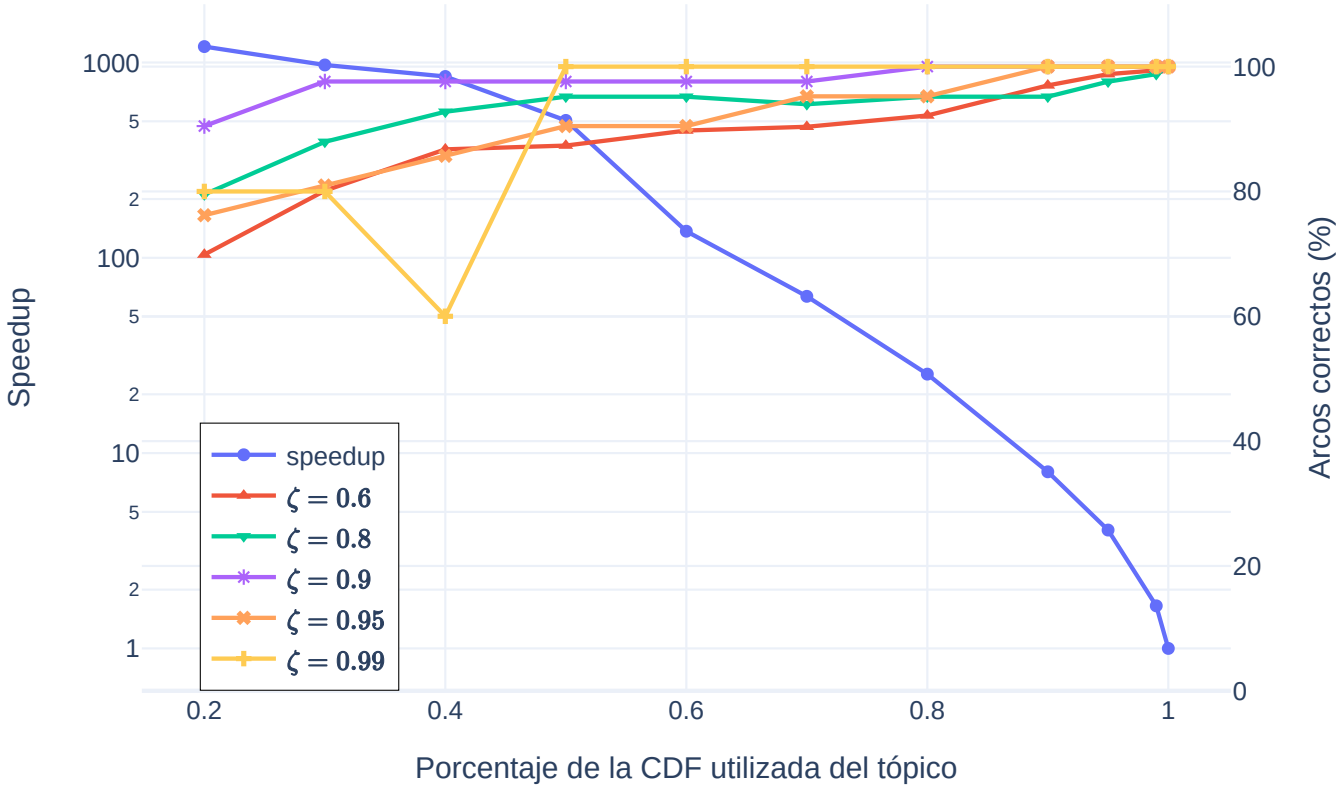


Figura 14: *Speedup* y porcentaje de arcos correctos al utilizar un menor porcentaje de la cdf de los tópicos en la construcción del grafo de similitud. El error de aproximación de la heurística es mostrado para diferentes puntos operantes ζ utilizados para podar el grafo completo.

Caso de estudio: Evolución del robo no presencial

El tópico presenta una tendencia a la baja en su participación, en el año 2011 alrededor del 44 % *tokens* provienen de este tópico y en el año 2016 se tiene un 32 %.

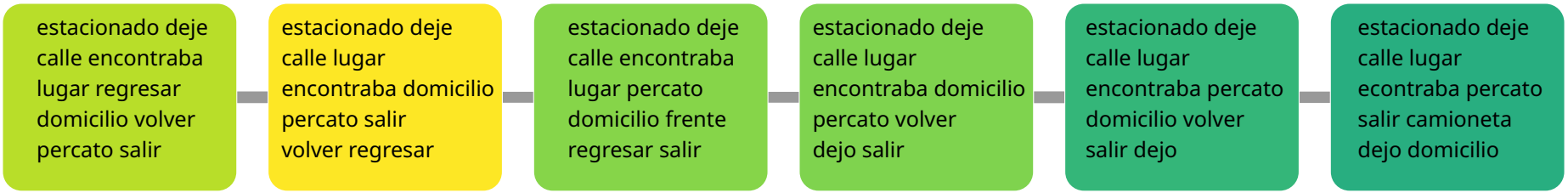


Figura 15: Evolución del tópico de robo de vehículo no presencial. El eje horizontal denota el tiempo en años, partiendo en el 2011 hasta el 2016. Mientras más claro el color del tópico más popularidad posee en su correspondiente época y mientras mayor es el grosor del arco entre dos tópicos mayor es su similitud.

Caso de estudio: Evolución del robo con violencia

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 %.

Figura 16: 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 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.

Conclusiones

- Se logra desarrollar una metodología exitosa para el descubrimiento de tópicos en el tiempo que permite capturar **dinamismos** tales como nacimiento, muerte, evolución, división y fusión.
- En contraste a trabajos anteriores, se incorpora WMD como medida de similitud permitiendo comparar tópicos que no poseen un vocabulario común.
- El análisis empírico del *trade off* entre precisión y *speedup* muestra que valores iguales o superiores $q = 0.6$ parecen bastante razonables.
- El análisis de sensibilidad del punto operante de la cdf muestra a que mayor valor mayor es el ratio de nacimiento/muerte, sin embargo, menor es el ratio división/fusión.
- Se identifica claramente la evolución del tópico robo no presencial, robo con violencia y el fenómeno del portonazo.

Trabajo futuro

- Una futura línea de investigación es el desarrollo de un modelo de tópicos dinámico basado completamente en redes neuronales, de esta forma la comparación entre tópicos de épocas adyacentes a través de sus *word embeddings se vuelve más natural.* En [Dieng et al., 2019] se propone una modelo de tópicos dinámico basado en redes neuronales, no obstante, mantiene fijo el número de tópicos en el tiempo.
- Extender la metodología a inferencia *online* sin alterar la estructura del grafo previo, debido a que cada vez que se incorpora una nueva época se vuelve necesario actualizar el úmbral de corte empírico.

Referencias I

Ahmed, A. and Xing, E. P. (2012).

Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream.

arXiv preprint arXiv:1203.3463.

Beykikhoshk, A., Arandjelović, O., Phung, D., and Venkatesh, S. (2018).

Discovering topic structures of a temporally evolving document corpus.

Knowledge and Information Systems, 55(3):599–632.

Blei, D. M. (2012).

Probabilistic topic models.

Communications of the ACM, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2006).

Dynamic topic models.

In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Referencias II

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).

Latent dirichlet allocation.

Journal of machine Learning research, 3(Jan):993–1022.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2019).

The dynamic embedded topic model.

arXiv preprint arXiv:1907.05545.

Dumais, S. T. (2004).

Latent semantic analysis.

Annual review of information science and technology, 38(1):188–230.

Escobar, M. D. and West, M. (1995).

Bayesian density estimation and inference using mixtures.

Journal of the american statistical association, 90(430):577–588.

Referencias III

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015).

From word embeddings to document distances.

In *International conference on machine learning*, pages 957–966.

Niculae, V. (2015).

Word mover's distance in python.

Pele, O. and Werman, M. (2009).

Fast and robust earth mover's distances.

In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012).

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.

Referencias IV

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005).

Sharing clusters among related groups: Hierarchical dirichlet processes.

In *Advances in neural information processing systems*, pages 1385–1392.

Wang, X. and McCallum, A. (2006).

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.

Wilson, A. T. and Robinson, D. G. (2011).

Tracking topic birth and death in lda.

Sandia National Laboratories.

Xu, W., Liu, X., and Gong, Y. (2003).

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.