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
Miembros de la comisión: Giorgiogiulio Parra, Ángel Jiménez

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

11 de octubre de 2021

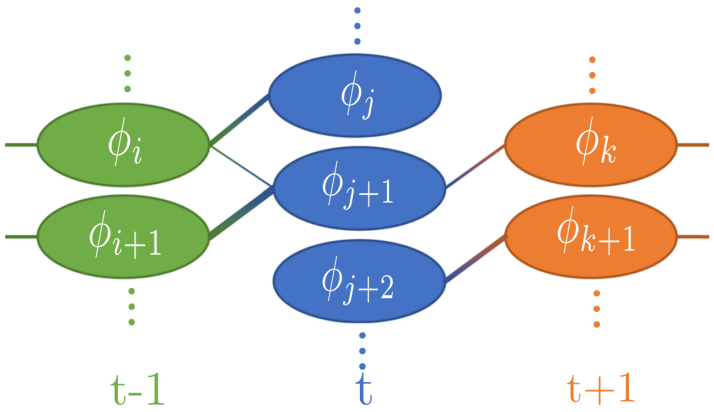
Contenidos

Motivación

2 EB de data por día

80 % no estructurado

12 % es analizada

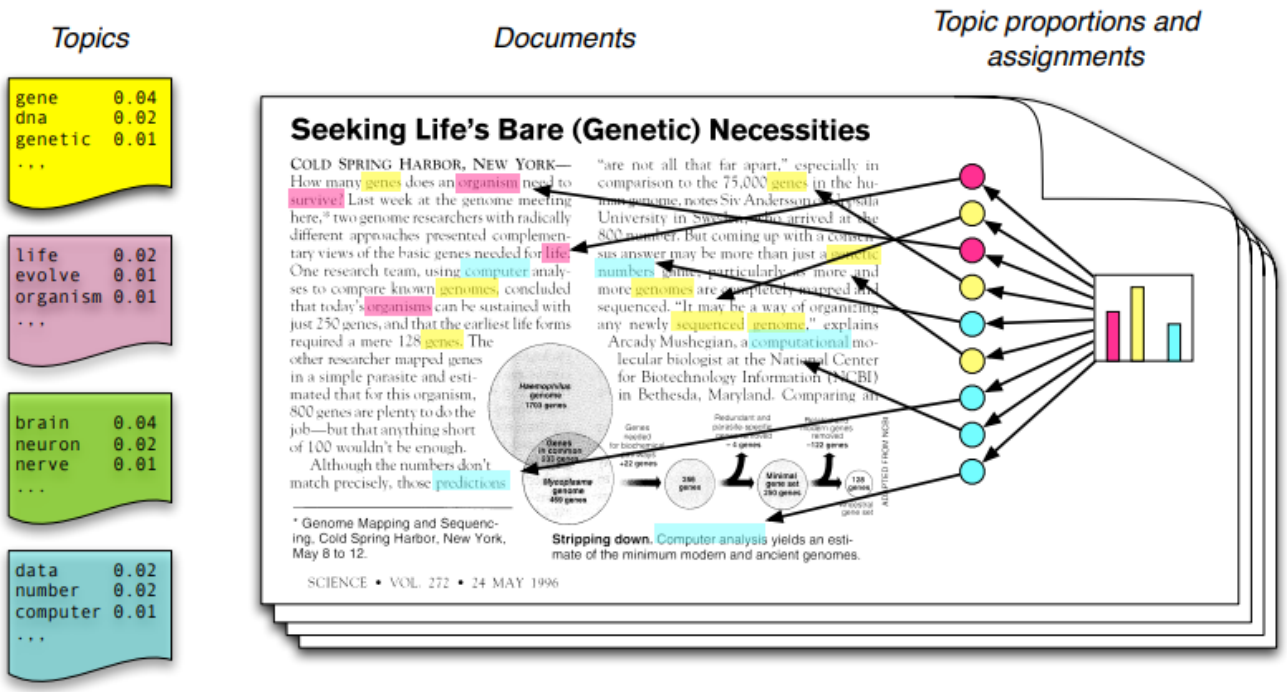


- Aumento del volumenes de textos requiere métodos automáticos de procesamientos.
- El modelamiento de tópicos dinámico permite descubrir nuevos fenómenos de base. Ejemplo: descubrimiento de nuevos tipos de accidentes de trayecto apartir de relatos de accidentes laborales.
- Este trabajo es una propuesta de modelamiento dinámico de nacimiento, muerte, evolución, división y fusión de tópicos.

https://www.sigmacomputing.com/blog/top-20-big-data-statistics/

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

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.



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 dinámismos:

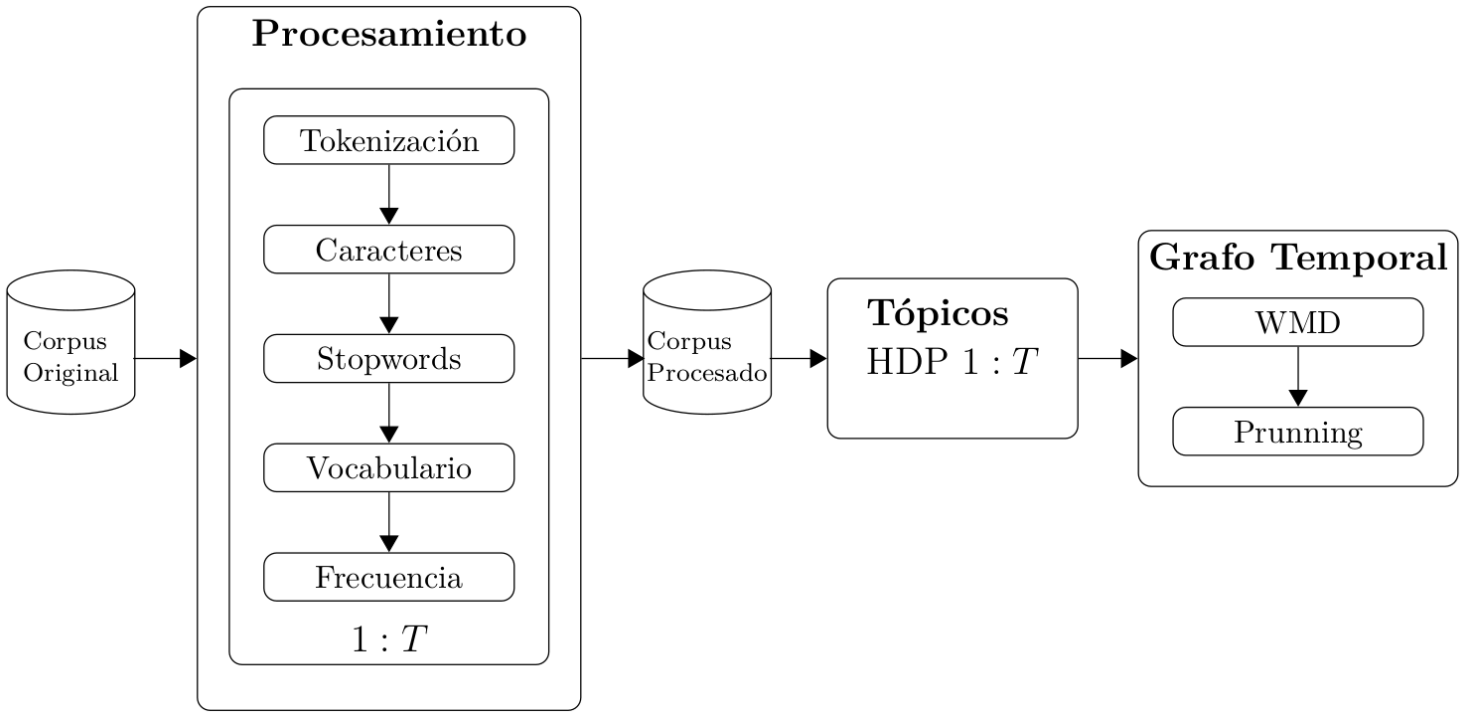
1. **Evolución de tópicos.**
2. **Dinámismo en la mezcla de tópicos.**
3. **Nacimiento, muerte, fusión y división de tópicos.**

Dentro de los modelos de tópicos dinamicos se tiene:

- **DTM** (Dynamic Topic Modelling)[Blei and Lafferty, 2006] y **TOC** (Topic Over Time)**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

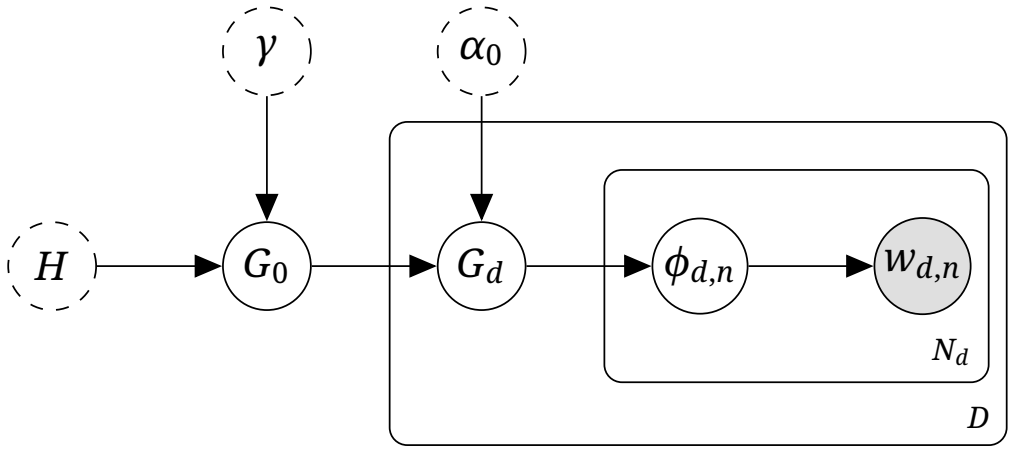
- División del corpus en épocas siendo cada época procesada mediante tokenización, eliminación de caracteres, eliminación de *stopwords*, filtro por vocabulario y filtro por frecuencia.
- Aplicación de HDP en cada época de manera independiente.
- Construcción del grafo computando la similitud WMD entre tópicos de épocas adyacentes y eliminación de arcos cuya similitud es menor al cuantil ζ de la distribución acumulada de la similitud.



Metodología propuesta: Hierarchical Dirichlet Process

HDP (Hierarchical Dirichlet Process) 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.

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

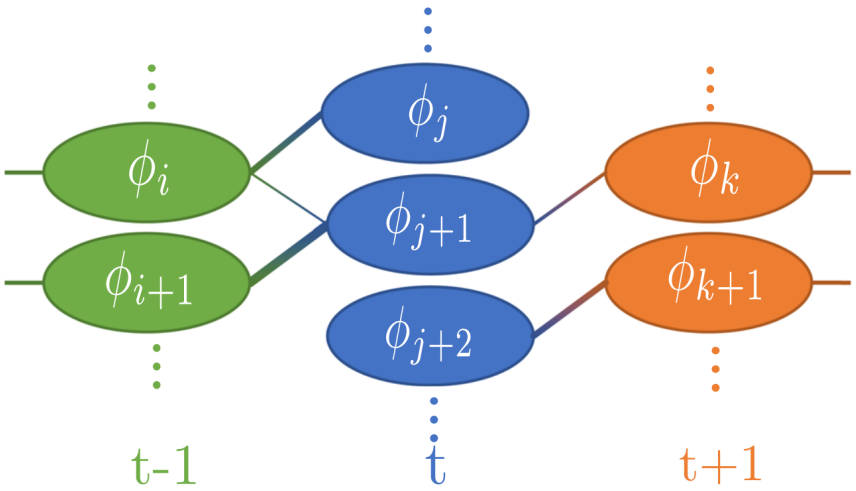


La discretitud de G_0 asegura:

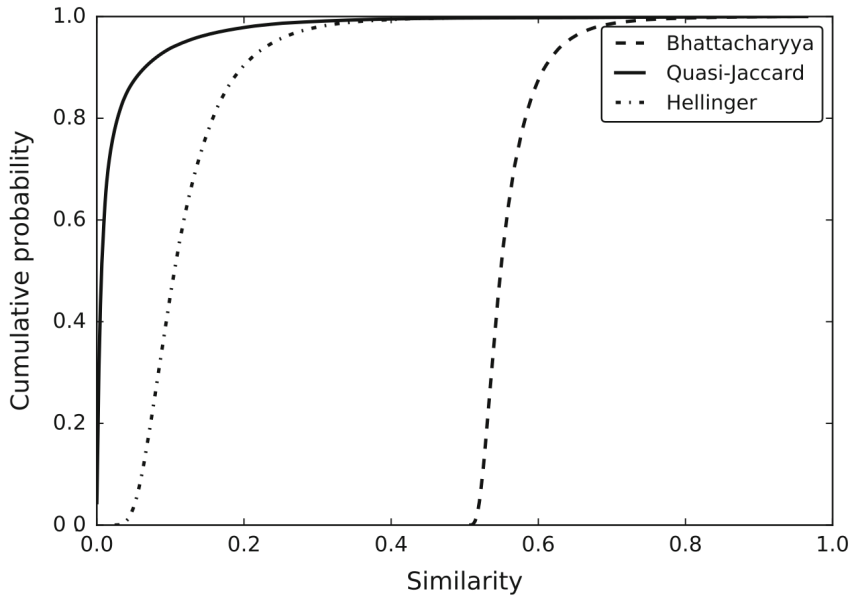
- A nivel corpus los documentos comparten el mismo conjunto de tópicos (*mixture components*).
- A nivel documento G_d hereda los tópicos de G_0, pero los pesos de cada tópico (*mixture proportions*) es específica del documento.

Metodología propuesta: 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 $\zeta \in [0, 1]$, reteniendo solo aquellas conexiones que cumplen $\rho(\phi_{t,i}, \phi_{t+1,j}) \leq \zeta$.



- 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.



Metodología propuesta: Word Mover's Distance

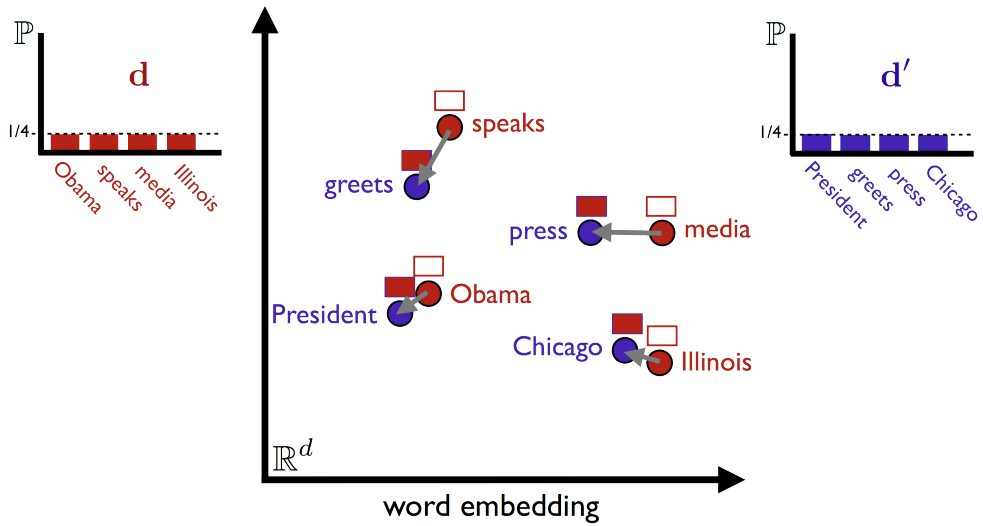
- Se escoge **WMD** (Word Mover's Distance) [Kusner et al., 2015]: distancia que permite comparar vectores sin vocabulario común ya que trabaja sobre el espacio de los*word embeddings*.
- Sea V_i y V_j los vocabularios del tópico i y j respectivamente, luego su WMD viene dado por $WMD(\phi_i, \phi_j)$:

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

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

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

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



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.

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 del vocabulario entre 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 $N \in [5, 30]$, 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.

Metodología propuesta: Configuracion de hiperparametros

HDP cuenta con tres hiperparámertos:

- 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.

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

Dynamic topic models.

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

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

Latent dirichlet allocation.

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

Referencias II

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.

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.

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.

Referencias III

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.

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.

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

Tracking topic birth and death in lda.

Sandia National Laboratories.

Referencias IV

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.