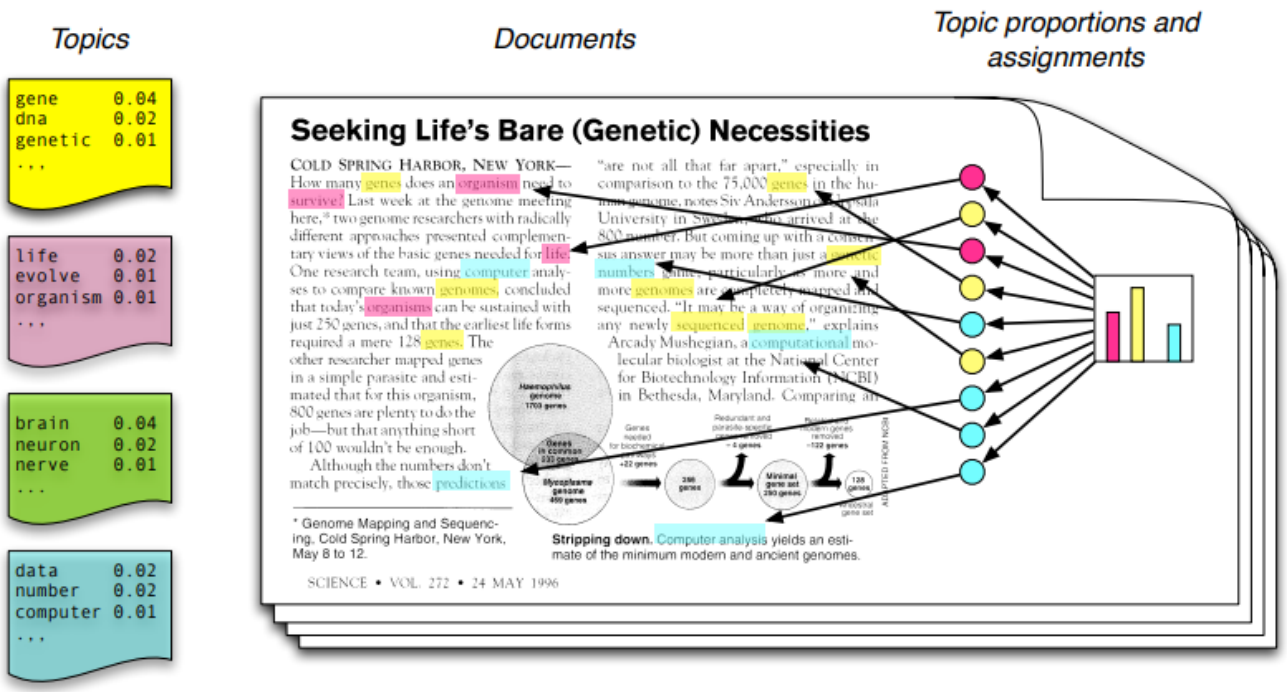
Contenidos

Motivación

- A medida que capturamos mas información se vuelve más difícil encontrar y descubrir lo que necesitamos. Volviéndose clave contar con herramientas que ayuden a organizar, buscar y entender grandes colecciones de datos.
- El modelamiento de tópicos permite enfocar la búsqueda en temas específicos. Por ejemplo, descubrir nuevas tedencias de investigación, analizar la evolución de la contigencia social en redes sociales, 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.

Revisión del estado del arte: Enfoque

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]. LDA necesita de antemano fijar el número de tópicos a descubrir y HDP lo infiere a partir del corpus.

Este trabajo se aborda 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].

Revisión del estado del arte: Modelamiento dinámico

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

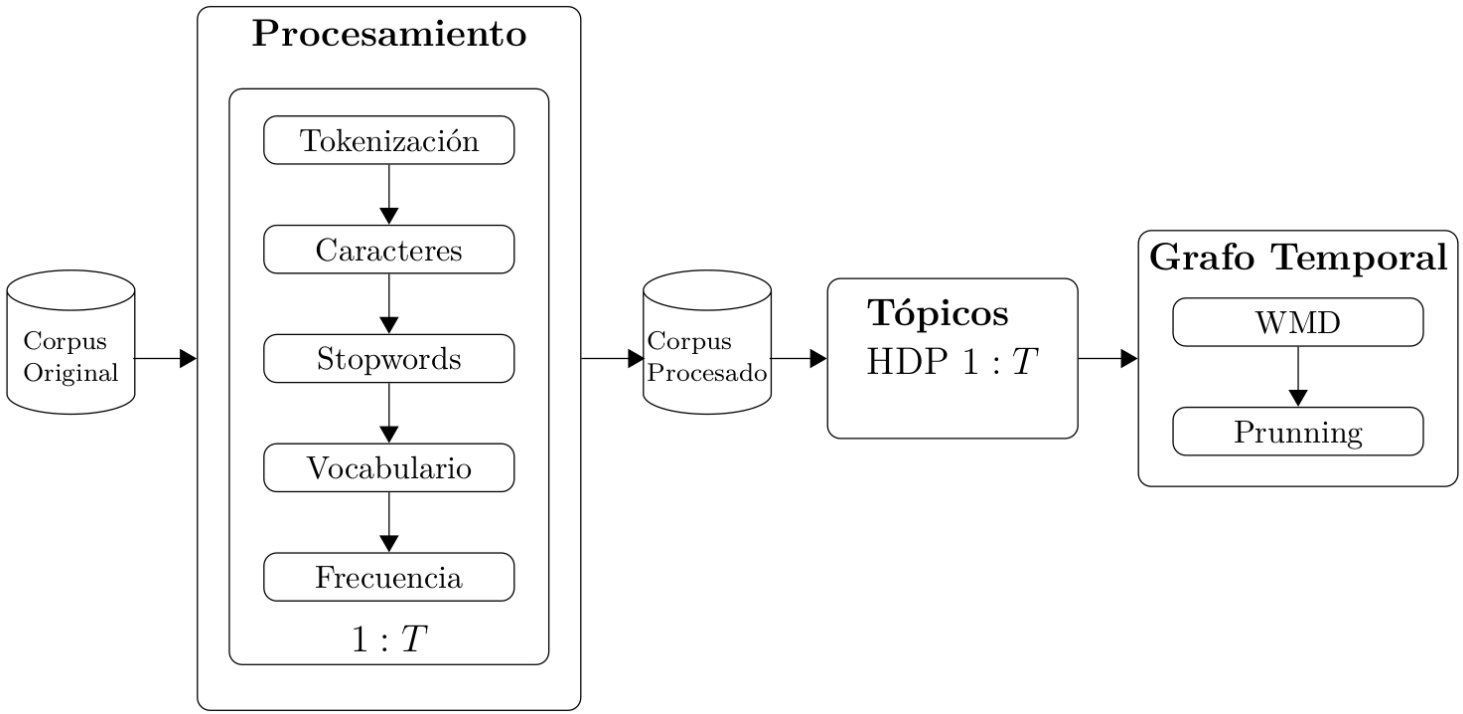
1. **Evolución de tópicos.**
2. **Dinámismo en la mezcla de tópicos.**
3. **Nacimiento, muerte, fusión y división de tópicos.**

Dentro de los modelos de tópicos dinamicos se tiene:

- Dynamic Topic Modelling (DTM) y Topic Over Time (TOC)[Wang and McCallum, 2006] permiten capturar el punto 1 y 2 manteniendo fijo el número de topicos en el tiempo.
- Dynamic Hierarchical Dirichlet Process (DHDP)[Ahmed and Xing, 2012] captura los tres puntos, con excepcion de fusión y división. No es una tecnología ampliamente usada y no cuenta con una implementación disponible.
- En [Wilson and Robinson, 2011] y [Beykikhoshk et al., 2018] se propone una metodología que permite capturar los dinámismos mencionados dividiendo el corpus en épocas, entrenar de forma independiente un modelo de tópico en cada época (LDA y HDP respectivamente), para finalmente unir los resultados obtenidos.

Metodología propuesta

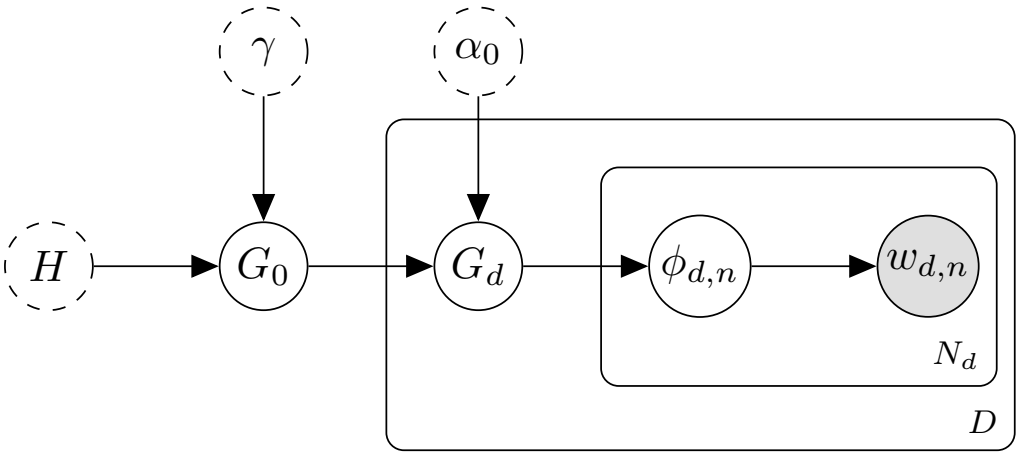
- Se divide el corpus original en épocas y a cada época se le aplican las siguientes cinco etapas en forma secuencial: tokenización, procesamiento de caracteres, eliminación de *stopwords*, filtro por vocabulario y filtro por frecuencia.
- Se aplica HDP de forma independiente sobre cada una de las épocas.
- Se construye el grafo temporal utilizando similitud WMD entre tópicos de épocas adyacentes. Finalmente, se podan los arcos cuya similitud es menor al cuantil ζ de la distribución acumulada del grafo *fully connected*.



Metodología propuesta: Hierarchical Dirichlet Process

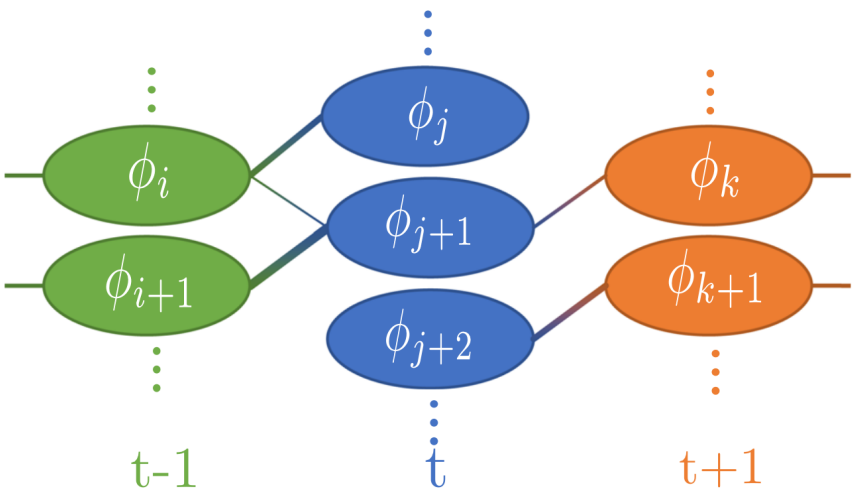
Hierarchical Dirichlet Process (HDP) es un *prior* jerárquico no paramétrico, el cual está formado por un DP cuya medida base G_0 es dibujada a partir de un DP. En el caso de modelamiento de tópicos, se tiene un medida global G_0 a nivel corpus que es dibujada a partir de un DP con medida base Dirichlet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

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

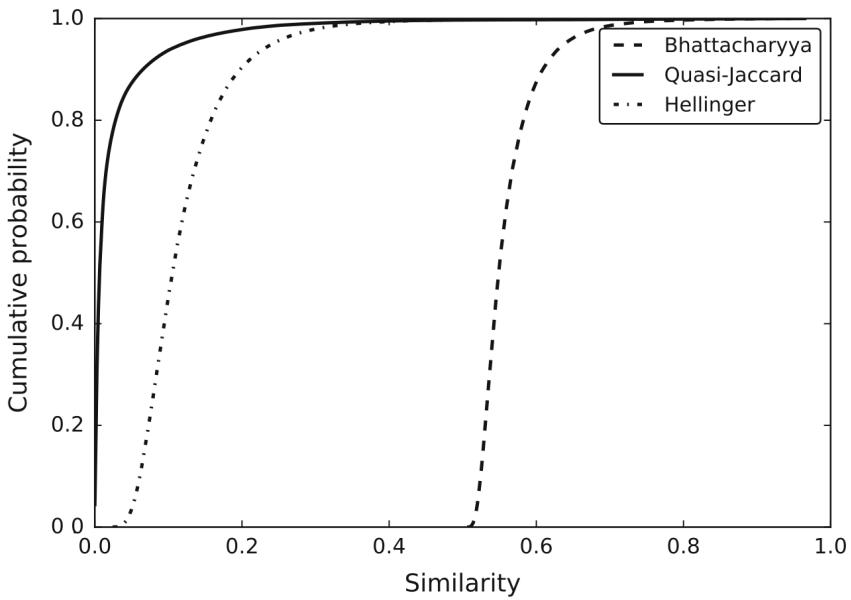


Metodología propuesta: Grafo de similitud temporal

- 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

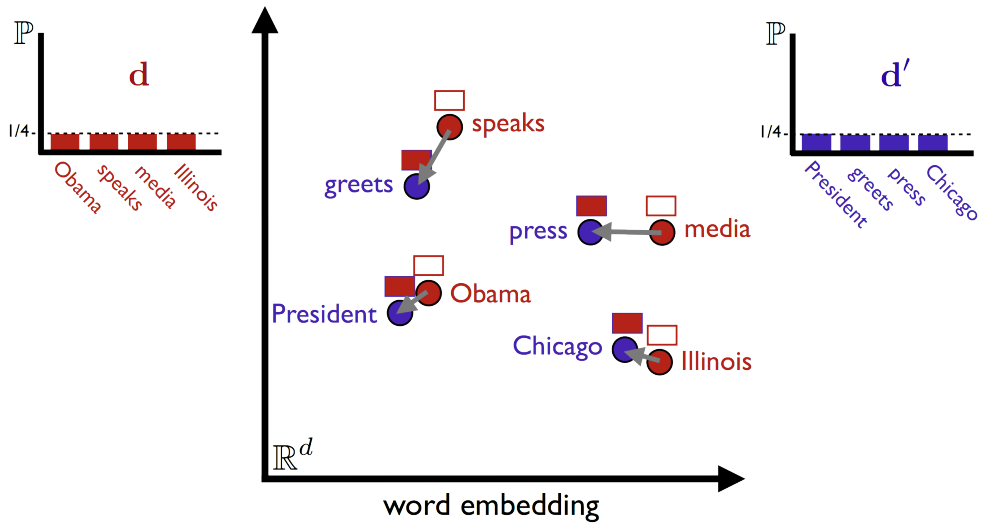
Las medidas de similitud suelen comparan vectores con mismo dominio y dimensión, por ende tópicos de épocas adyacentes deben compartir un vocabulario común. Para lidiar con este problema se propone utilizar Word Mover's Distance (WMD)[Kusner et al., 2015], medida que trabaja sobre el espacio de los word 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.

- Sea N el tamaño del vocabulario entre dos épocas adyacentes.
- Sea $\{x|Ax = b, x \geq 0\}$ la región factible 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.
- Usando el algoritmo desarrollado por [Pele and Werman, 2009] se tiene que el mejor tiempo promedio escala $\mathcal{O}(N^2 logN)$.

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 $\eta \in (0, 1)$, esto generará 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. Valores razonables son $[0.8, 0.95]$, de esta manera se conserva el *core* de palabras del tópico y se disminuye de manera significativa el tiempo de cómputo.
- $\zeta \in [0, 1]$ cuantil de corte de la cdf de las similitudes del grafo fully connected. 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.

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

Latent dirichlet allocation.

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

Dumais, S. T. (2004).

Latent semantic analysis.

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

Referencias II

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.

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.

Referencias IV

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.