Contenidos

1. Motivación
2. Revisión del estado del arte
3. Metodología propuesta
4. Descubrimiento de tópicos en robo de vehículos
5. Conclusiones y trabajos futuros

Motivación

- A medida que capturamos mas informacion se vuelve mas dificil encontrar y descubrir lo que necesitamos. Siendo clave contar con herramientas que nos ayuden a organizar, buscar y entender grandes colecciones de datos.
- Con modelamiento de topicos podemos enfocar nuestra busqueda en temas especificos. Por ejemplo, descubrir nuevas tedencias de investigacion, analizar la evolucion de la contigencia social en redes sociales, estudiar la efectividad de campanas publicitarias en base a reviews, etc.

Motivación: Objetivo

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.

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.

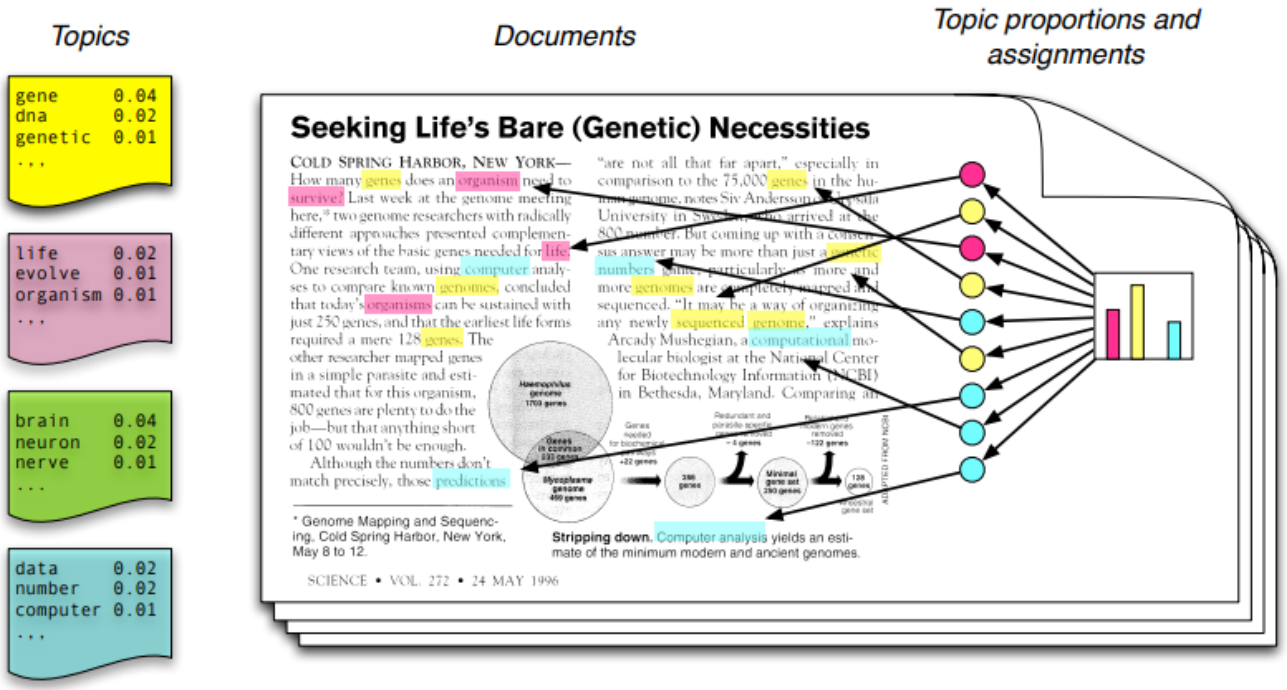

Figura 1: Ejemplo de topicos descubiertos usando LDA en un corpus de publicaciones científicas.

Revisión del estado del arte: Tipos de modelos de topicos

Las tecnicas de modelamiento de topicos suelen estar basadas en factorizacion matricial o en modelos probabilisticos generativos. A continuacion algunos ejemplos:

- 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.

Se escoge 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 dinamico

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 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: Resumen

La metodologia propuesta para el descubrimiento de topicos en el tiempo esta basada en (i) discretizacion del corpus en epocas, (ii) descubrimiento de topicos en cada epoca mediante Hierarchical Dirichlet Process (HDP), (iii) la construccion de un grafo de similitud entre topicos de epocas adyacentes, el cual permite modelar cambios entre los topicos como: nacimiento, muerte, evolucion, division y fusion. En contraste a trabajos anteriores, la metodologia propuesta utiliza Word Mover's Distance (WMD) como medida de similitud entre topicos, medida que destaca por ser robusta a topicos que no poseen un vocabularion comun, debido a que trabaja con sus word embeddings.

Metodología propuesta: Hierarchical Dirichlet Process

Metodología propuesta: Grafo de similitud temporal

Metodología propuesta: Word Mover's Distance

Metodología propuesta: Configuracion de hiperparametros

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.

Referencias II

Dumais, S. T. (2004).

Latent semantic analysis.

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

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.

Referencias III

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.