December 5, 2022

Abstract

1 Conclusiones de artículos leidos e ideas a implementar

1.1 Estrategia para la selección del dataset y su procesado

1.2 Estrategia de selección de modelos

Para elegir el modelo que más se adapta a este caso de uso se han generado tñopicos con diferentes técnicas utilizadas en la literatura para extraer los tópicos inherentes en la colección de textos. Los modelos generados han sido los siguientes:

- BERTopic
- DTM
- D-ETM

Asumimos que cada tópico debe representar un concepto específico y que cada concepto está definido como las 15 palabras con mayor probabilidad para ese tópico. Observamos la coherencia de los tópicos generados en los últimos 5 años para conocer si la evolución de los tópicos es homogénea y estable. De esta forma podemos saber si los modelos generan modelos que evolucionan progresivamente y representan correctamente la colección de textos para elegir el mejor. Nos basaremos concretamente en las métricas CV y UMass (interpretabilidad y especificidad respectivamente).

1.3 Análisis del modelo generado

Para etiquetar los diferentes tópicos hemos decidido tomar las palabras que han sido la top1 del tópico en cualquiera de los espacios temporales. Esta estrategia nos permite definir los tópicos mediante la o las palabras más importantes. Además, para poder medir y comparar la especificidad semántica de los tópicos se ha decidido utilizar la distancia semántica de las palabras de cada etiqueta. De esta forma, si un tópico es definido por una palabra, este tendría distancia cero y, por ende, sería una etiqueta perfecta. Mientras que, los tópicos que están definidos por una étiqueta con más de una palabra serán más plurales y serán menos específicos.

Sin embargo, mediante un estudio superficial de las etiquetas generadas se ha observado que muchas palabras son tendencias con alta relevancia en un espacio temporal y muy poco en los demás. Esto implica que los tópicos estan formados por tendencias. Por ello, la estrategia de etiquetado ha sido levemente modificada. Ahora, para que una palabra cuente como etiqueta, además de haber sido top1 en cualquier fecha ha de permanecer en el top5 palabras de forma estable durante todas las fechas. Además, si un tópico no cumple con palabras que satisfagan este requisito se asumirá que el tópico es demasiado volátil y no puede ser utilizado para estudiarlo.

Para contabilizar esto también hemos desarrollado una métrica de drop para ponderar la cantidad de tópicos volátiles respecto al total. De esta forma podemos comparar los diferentes modelos creados y cual genera tópicos más estables.

Ante los resultados obtenidos y los tópicos tan variados se ha decidido incrementar el número de tópicos para observar si de esta manera se consiguen tópicos más específicos ya que se conoce que esto genera tópicos más concretos(Hay varios trabajos que nos lo confirman y el nuestro también lo demuestra).

Observamos que DTM es el modelo que da mejor rendimiento, sin emabrgo entre las top 15 words de los tópicos no aparecen medicamentos. Además, observamos que aparecen el SARS, MERS y el SARS-CoV-2 en el mismo tópico, lo que impide que podamos relacionar los fármacos con cada enfermedad.

2 Artículos leidos

Gran parte de los artículos son de autores chinos y tratan del análisis de opiniones. Muchos con financiación estatal china. Además, es un denominador común en las investigaciones aplicar estas técnicas sobre el texto recogido de diferentes redes sociales.

[Yao and Wang(2020)]: Estos autores presentan la idea de **cada tópico tiene una trayectoria definida** como el camino que atraviesa el centroide del tópico a través de los diferentes espacios temporales. Además, definen que un tópico es estático o dinámico, es decir si cambia mucho o poco en el tiempo (porque el cambio es inevitable), en base a la distancia que recorre el tópico. Fijan una distancia máxima por fecha que vale para establecer si un tópico cambia mucho o poco.

[Lee and Krumm(2011)]: La trayectoria espacial la definen estos autores (cita del anterior artículo).

[Alazba et al.(2022)Alazba, Abouhagar, Al-Harbi, Al-Jamimi, Sultan, and Al-Zaidy]: Este artículo no me ha gustado nada, parece muy sesgado, simple, los modelos han sido elegidos a dedo y hay muchas cosas hechas a mano. Sin embargo he encontrado algunas cosas interesantes.

- Definen el tópico trending o dominante como aquel que predomina en la mayoría de artículos. Es decir, teniendo la distribución de tópicos documentos extraen el tóptico que predomina en cada documento, con esto cuentan cuál es el tópico más repetido y ése es.

- Definen los N tópicos más importantes del corpus como los N que más predominan (siguiendo el proceso anterior).

- Para etiquetar los tópicos utilizan el documento más relevante y las top 20 palabras y lo hacen a mano.Asignan nombre basandose en las palabras y lo confirman con el documento.
- Analizan la cantidad de artículos financiados y los tópicos más financiados.

2.1 Otros enfoques

[Churchill and Singh(2022)]: Estos autores presentaron un dynamic topic-noise discriminator. Proponen el Dynamic Noiseless Latent Dirichlet Allocation, adaptan el topic-noise model a un espacio temporal de las redes sociales, es una adaptación temporal del Noiseless Latent Dirichlet Allocation. Este enfoque no solo propone que un tópico evoluciona a lo largo del tiempo, sino que también asume que un tópico está formado por palbras y ruido. Creando una distribución de ruido. A los de este estudio D-ETM tampoco les genera tópicos de calidad. Estos analizan la evolución del tópico de la vacuna del coronavirus con el corpus del COVID-19 (NO DICE CUAL).

2.2 Colecciones de artículos

[Sleeman et al.(2021)Sleeman, Finin, and Halem]: Estos autores hacen un estudio muy parecido al que nosotros buscamos pero en otro contexto. Utilizan documentos técnicos sobre una colección de documentos de ciberseguridad. Además, presentan los documentos, los conceptos y los tópicos en gráfos de conocimiento para ayudar a la integración y poder hacer consultas. "'When we build topic models over time, topics evolve over time based on the documents in the collection at that time point. Our observation is as we increased the number of topics, we saw more granularity among topics. Topics represented more narrow mixtures. We also observed concepts that drop off of one topic and fall into another at various points in time".

2.3 Social Media Data

[Golino et al.(2022)Golino, Christensen, Moulder, Kim, and Boker]: Estos autores analizan los mensajes publicados en redes sociales que instigaron a la opinión pública a desconfiar de los procesos electorales de EEUU en el año 2016. No se pueden extraer cosas muy diferentes a las de otros artículos pero me ha gustado el concepto de que hay cuestiones que influencian la opinión pública. Hacen un análisis de opinión que actualmente no nos interesa pero está bien saber que herramientas han utilizado y como, sobre todo porque los tweets no contienen mucho texto. De hecho, podemos analizar la largura media de un abstract para compararla con un tweet y explicar que puede haber modelos que trabajen mejor con volumenes de datos de menor longitud. En la sección de análisis de datos de diferentes formas (citan algunos enfoques). Crea un modelo de 10 tópicos que es bastante gráfico.

[Ghoorchian and Sahlgren(2020)]: Estos autores crean una solución para aplicar DTM en textos cortos de redes sociales. En teoría utilizan grafos de conocimiento pero como por ahora esto no nos interesa no he profundizado. Sin embargo esta guay la definición que hacen de las técnicas de modelado de tópicos: the goal is to reduce the high-dimensional space of words into a significantly low-dimensional and semantically rich space of topics (citan un artículo).

[Tabassum et al.(2021)Tabassum, Gama, Azevedo, Teixeira, Martins, and Martins]: En este artículo usan los modelos de tópicos para extraer tópicos de los tweets. Sin embargo, los tópicos están definidos por los hastags y los utilizan para definir los tópicos que se van a tratar.

References

[Yao and Wang(2020)] Fang Yao and Yan Wang. Tracking urban geo-topics based on dynamic topic model. *Computers, Environment and Urban Systems*, 79:101419, 1 2020. ISSN 0198-9715. doi: 10.1016/J.COMPENVURBSYS.2019.101419.

[Lee and Krumm(2011)] Wang-Chien Lee and John Krumm. Trajectory preprocessing, 2011.

[Alazba et al.(2022)Alazba, Abouhagar, Al-Harbi, Al-Jamimi, Sultan, and Al-Zaidy] Amal Alazba, Leina Abouhagar, Randah Al-Harbi, Hamdi A. Al-Jamimi, Abdullah Sultan, and Rabah A. Al-Zaidy. Detection of research trends using dynamic topic modeling. 2022. doi: 10.1109/CDMA54072.2022.00031.

[Churchill and Singh(2022)] Rob Churchill and Lisa Singh. Dynamic topic-noise models for social media, 2022.

[Sleeman et al.(2021)Sleeman, Finin, and Halem] Jennifer Sleeman, Tim Finin, and Milton Halem. Understanding cybersecurity threat trends through dynamic topic modeling. *Frontiers in Big Data*, 4, 6 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.601529.

[Golino et al.(2022)Golino, Christensen, Moulder, Kim, and Boker] Hudson Golino, Alexander P. Christensen, Robert Moulder, Seohyun Kim, and Steven M. Boker. Modeling latent topics in social media using dynamic exploratory graph analysis: The case of the right-wing and left-wing trolls in the 2016 us elections. *Psychometrika*, 87:156–187, 3 2022. ISSN 0033-3123. doi: 10.1007/s11336-021-09820-y.

[Ghoorchian and Sahlgren(2020)] Kambiz Ghoorchian and Magnus Sahlgren. Gdtm: Graph-based dynamic topic models. *Progress in Artificial Intelligence*, 9(3):195–207, Sep 2020. ISSN 2192-6360. doi: 10.1007/s13748-020-00206-2. URL https://doi.org/10.1007/s13748-020-00206-2.

[Tabassum et al.(2021)Tabassum, Gama, Azevedo, Teixeira, Martins, and Martins] Shazia Tabassum, João Gama, Paulo Azevedo, Luis Teixeira, Carlos Martins, and Andre Martins. Dynamic topic modeling using social network analytics, 2021.