Técnicas matriciales para la clasificación de discursos presidenciales

Ian Bounos
Universidad de Buenos Aires, Argentina
bounosian@gmail.com

En este trabajo se muestra cómo pueden utilizarse métodos de reducción de dimensionalidad basados en matrices para la clasificación de autores de discursos presidenciales. La representación de textos como matrices de frecuencias, en la cual cada columna es una palabra del vocabulario, suele presentar el desafío de la alta dimensionalidad, por lo cual es preciso utilizar técnicas para reducir dicha dimensión. En este estudio, se emplean 1108 discursos de los presidentes Alberto Fernández, Cristina Fernández de Kirchner y Mauricio Macri, obtenidos mediante técnicas de scraping de páginas oficiales. Se utilizan dos métodos de reducción de dimensionalidad basados en matrices: el Análisis de Componentes Principales (PCA) y la Factorización No Negativa de Matrices (NMF), con el objetivo de reducir la dimensión de las matrices y, en primer lugar, obtener una visualización de los discursos. En una segunda instancia, se utiliza la versión reducida como entrada para un algoritmo de K vecinos más cercanos, con el fin de clasificar los textos, es decir, determinar a qué presidente corresponde cada uno, con una separación entre el conjunto de datos de entrenamiento y testeo. Se concluye con una comparación de ambos métodos, no solo en términos cuantitativos, evaluando su rendimiento predictivo, sino también en términos cualitativos para permitir una interpretación más profunda de los resultados obtenidos.

Trabajo en conjunto con Dirección de Juan Pablo Pinasco (Universidad de Buenos Aires).

Referencias

[1] Natural language processing

[2] Ciencia de datos

[3] Non negative matrix factorization

[4] Análisis de componentes principales