Metodología

En el siguiente apartado buscaremos exponer cuáles fueron los pasos a seguir en nuestro trabajo,describiremos los insumos empleados, las herramientas utilizadas y también los fundamentos teóricos de las mismas.

En primer lugar, nuestro insumos fueron las notas periodísticas del NY Times y del Ny Post, publicadas en sus respectivos portales web, con un rango temporal que va desde marzo hasta octubre del 2020. Para acceder al contenido de las mismas utilizamos diferentes librerías de Python (requests, BeautifulSoup). A su vez, utilizamos otras tantas para lograr la correcta manipulación de los datos obtenidos(Pandas, numpy, re). Una vez extraída la información, realizamos un preprocesado de los textos, las fechas y demás componentes extraídos, para así luego poder trabajar correctamente con las herramientas posteriores.

Ahora que ya hemos conseguido los insumos y los tenemos preprocesados, pasamos a la parte del análisis de tópicos. Para esta tarea utilizaremos el paquete “BERTopic”. El mismo nos permitirá acceder a aquellos tópicos que se encuentran con mayor frecuencia en los diferentes textos a analizar. Para decirlo de una forma sencilla, podremos encontrar aquellos ejes temáticos que suelen repetirse con frecuencia en las noticias con las que trabajamos.

¿De qué forma es que actúa “BERTopic”? Básicamente es una técnica de modelado de tópicos que se basa en lo último de lo último en lo que respecta a procesamiento de lenguaje natural. El algortimo utiliza Bert para poder lograr representaciones densas (embeddings) de los textos, reduce la dimensionalidad de estas representaciones con la herramienta UMAP, las agrupa según su similaridad con HDBSCAN, y luego encuentra los términos más representativos de cada uno de los grupos creados baśandose en la métrica c-TF-IDF , que es una variación de la clásica TF-IDF (*Term frequency – Inverse document frequency)* *.* Así es que se logra encontrar tópicos fácilmente interpretables manteniendo las palabras más importantes en las descripciones de esos tópicos.

Pero, ¿Qué son estas cosas que mencionamos en el párrafo anterior? Los embedding son representaciones vectoriales de los términos presentes en los textos que tienen un significado que se deriva de las palabras que le hacen de contexto a cada término. Decimos que son representaciones densas ya que el embedding de cada palabra conserva la semántica propia de ella. A su vez, podemos construir embeddings de un texto entero a partir de realizar operaciones matemáticas con los embeddings de las palabras que conforman el texto. UMAP es un algoritmo de reducción de dimensionalidad mientras que HDBSCAN es un algoritmo de clusterización, ambos funcionan de manera óptima en conjunto, dadas las particularidades matemáticas de los mismos. Por otra parte, la métrica TF-IDF es una medida numérica que nos permite cuantificar qué tan relevante es una palabra en un texto o grupo de textos. La variación c-TF-IDF agrupa todos los textos pertenecientes a cada categoría creada previamente, como si fuesen un sólo y gran texto, y les aplica TF-IDF para encontrar las palabras más relevantes en cada tópico.