Reduciendo la dimensionalidad de los textos

La reducción de la dimensionalidad es un aspecto central en las aplicaciones de procesamiento de textos, ya que permite determinar el conjunto de características que describe mejor los documentos. Además, es esencial para mejorar el rendimiento de los algoritmos de aprendizaje.

¿Qué caminos podemos seguir para realizar esta reducción de la dimensionalidad? Algunos son:

- Reducir el diccionario de los datos. Cuando se determina el lexicón de los textos hay que tener en cuenta que este puede contener muchas palabras, las cuales pueden no ser todas necesarias para la tarea que se quiere resolver. Una forma de especializar este diccionario es incorporar sólo las palabras más frecuentes o términos cuyas frecuencias superan un cierto umbral. También se podrían incorporar sólo los términos de las clases de interés.

- Métodos de selección de características. Estos métodos evalúan la importancia de cada atributo y seleccionan un subconjunto de características que se consideran más relevantes. Pueden emplearse técnicas filtros o métodos basados en modelos.

- Métodos de extracción de características. Se pueden aplicar técnicas de reducción de la dimensionalidad por transformación a las representaciones vectoriales de documentos, como el análisis de componentes principales (PCA), para construir un espacio de características reducido.

- Incrustaciones de palabras. Otra forma de reducir la dimensionalidad y mejorar la representación de los datos de texto es utilizar incrustaciones de palabras (*word embedding*), que son vectores densos y de baja dimensión que capturan las relaciones semánticas entre las palabras. Las incrustaciones se aprenden de grandes cantidades de datos de texto utilizando, modelos de redes neuronales, como Word2Vec, que explotan el contexto de las palabras; o mediante una matriz de coocurrencia que captura la frecuencia con la que aparecen juntas las palabras en el corpus, que es el caso de Glove. También disponemos de los *embeddings*

contextuales pre-entrenados, como BERT (*Bidirectional Encoder Representations from Transformers*), que capturan relaciones semánticas y contextuales de una palabra teniendo en cuenta tanto las palabras anteriores como las posteriores en una oración o incluso en el documento completo.

Bibliografía

Vajjala, S., Majumder, B., Gupta, A., Surana, H. (2020). *Practical Natural Laguage Processing*. O'Reilly Media, Inc.

Tunstall, L., von Werra, L., Wolf, T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.