Informe TPº1 NLP

Tecnicatura Universitaria en Inteligencia Artificial

Integrantes:

- Mateo Rovere
- Valentín Dalmau

Profesores:

- Juan Pablo Manson
- Andrea Cavallo
- Alan Geary
- Ariel D'Alessandro

Ejercicio 1

Para construir el dataset buscamos blogs de 4 temas distintos (cocina, ciencia, salud y finanzas), el hecho de que fueran blogs facilitó mucho el webscrapping porque todos los títulos tenían el mismo formato, y en general los párrafos finales eran los de menos interés porque podían hablar de temas específicos del blog en vez de la temática elegida. Para realizar el dataset creamos 4 funciones similares que permitían hacer un webscrapping a cada blog particular, la diferencia principal en cada función es la clase del título. En estas funciones conseguimos los datos de cada página con el método get de la librería requests y con Soup conseguimos el texto. Finalmente creamos una fila del dataframe con la url, título, texto y categoría correspondientes. Luego llamamos a cada función 10 veces en cada url (guardados en listas para facilitar el acceso) y terminamos de crear el dataframe. Para no contaminar la muestra de texto, eliminamos la palabra Ribera del dataframe porque aparecía mucho en el blog de salud (era el nombre del blog).

Ejercicio 2

Para resolver este ejercicio utilizamos 2 métodos distintos, Tf-idf y Bert. Como era de esperarse los resultados con Bert fueron muy superiores ya que es un modelo mucho más complejo. Además de tener una precisión de regresión logística mucho mayor pudimos crear un predictor de frases solo con los títulos de cada categoría, que tuvo una precisión perfecta incluso con frases no triviales.

Ejercicio 3

Lo primero que hicimos fue limpiar y normalizar el texto, lematizando las oraciones y eliminando las stop words y signos de puntuación. Luego calculamos la frecuencia de las palabras en los textos de cada categoría e hicimos una nube de palabras utilizando la librería WordCloud. La primera vez que lo hicimos notamos que la palabra Ribera aparecía en la categoría salud y por eso la eliminamos más adelante. Notamos que prácticamente todas las palabras en la nube parecían tener una relación directa con la categoría (salvo tal vez “importante” en finanzas, que será una palabra que usan demasiado en ese blog). Por otro lado también notamos que algunas palabras se repiten entre salud y ciencia, lo cuál tiene sentido porque son categorías que tienen puntos en común, particularmente viendo la nube de palabras de ciencia, parece ser que los artículos elegidos fueron mayormente relacionados con la biología.
En general las palabras más usadas de cada categoría parecen reflejar con precisión la temática pertinente, por lo tanto este análisis puede ser útil para proporcionar palabras claves correspondientes a cada categoría.

Ejercicio 4

Este ejercicio también lo realizamos de 2 formas distintas, con la librería sentence_transformers y con Doc2Vec. Como esperábamos, la primera alternativa dio los mejores resultados, siendo que era un modelo más complejo.
En el caso de sentence transformers, encontramos un resultado muy bueno, ya que el modelo pudo comprender con claridad el significado semántico de las oraciones. Por ejemplo, puede detectar la similitud entre “diagnóstico” y “detectar” o entre “tumores” y “cáncer”. El segundo modelo dio resultados significativamente peores, donde realmente no pudimos encontrar una relación entre los títulos que el modelo consideró más parecidos entre sí.

Ejercicio 5

Para este ejercicio usamos un resumen abstractivo, dado que intentamos con un modelo extractivo pero los resultados no eran esperados, y de la manera de abstractivo el resumen era sorprendentemente bueno, hasta hay veces que se parecía al titulo del articulo