Diplomado - Inteligencia Artificial y Ciencia de Datos

Vocabulary Challenge - Mauricio Miranda Baron

Proceso

Lo primero que hice fue descargar el libro *Los Miserables* de Victor Hugo desde la página de la Fundación Carlos Slim (https://aprende.org/pruebat?sectionId=6).

Para convertir el archivo PDF a formato CSV, utilicé la plataforma *iLovePDF* (https://www.ilovepdf.com/es), que permite la conversión entre distintos tipos de archivos.

Una vez obtenido el archivo en formato CSV, comencé el análisis en Python, ya que ofrece diversas funciones para manipular cadenas de texto y considero que es más fácil de usar. Dado el tamaño del archivo, la velocidad de procesamiento de Python no representaba un problema. Para el análisis, utilicé la biblioteca *pandas*, además de *string* y *re*, que proporcionan herramientas para trabajar con cadenas de texto y expresiones regulares, respectivamente.

Inicié con un análisis exploratorio del archivo CSV y noté la presencia de caracteres especiales, espacios en blanco y saltos de línea innecesarios. Para limpiarlo, creé un diccionario con diversos caracteres y signos especiales, aprovechando los que ya proporciona Python. Luego, empleé una expresión regular para eliminar caracteres especiales, números y símbolos presentes en el libro, pero que Python no detecta automáticamente. También eliminé espacios en blanco innecesarios y convertí todas las palabras a minúsculas.

Durante este proceso, recorrí el archivo y, conforme limpiaba los datos, recolectaba las palabras. Implementé un pequeño algoritmo para almacenarlas en un diccionario, contando la frecuencia con la que aparecían en el texto.

Al finalizar, ordené las palabras tanto de menor a mayor como de mayor a menor frecuencia y mostré las 100 palabras más y menos repetidas. Finalmente, convertí el diccionario en un *DataFrame* de *pandas* y lo guardé en formato *Parquet* para facilitar su almacenamiento y análisis posterior.

```
Número total de palabras: 109280
```

```
Tamaño del diccionario: 13312 palabras
```

```
100 palabras menos repetidas
dominio => 1
prohibida => 1
ajenos => 1
zúrich => 1
carso => 1
ampliación => 1
granada => 1
c => 1
contactopruebatorg => 1
charlesfrançoisbienvenu => 1
interesa => 1
exactos => 1
circulado => 1
ocupa => 1
reservándole => 1
consagrada => 1
galanterías => 1
sobrevino => 1
precipitáronse => 1
diezmadas => 1
perseguidas => 1
acosadas => 1
emigró => 1
trágicos => 1
...
memorable => 1
verdaderas => 1
celdas => 1
renueva => 1
```

```
100 palabras mas repetidas
de => 5323
la => 3918
que => 3504
y => 3122
el => 3081
en => 2835
a => 2489
se => 1633
un => 1601
no => 1498
los => 1353
una => 1316
su => 1245
por => 936
las => 935
con => 924
había => 858
del => 813
al => 756
es => 749
lo => 717
le => 667
era => 650
más => 510
...
decir => 112
voz => 111
allí => 107
ojos => 107
```

Experiencia

En mi opinión, fue un buen ejercicio. A simple vista puede parecer sencillo, pero al implementarlo te das cuenta de los retos que implica, especialmente la limpieza del texto. En este caso, al tratarse de un libro, se puede confiar en que la mayoría de las palabras están bien escritas y tienen una ortografía correcta.

Actualmente, estoy desarrollando mi proyecto terminal en la UAM-I junto con un grupo de estudiantes formado por un profesor. Nuestro equipo se enfoca en Ciencia de Datos y, en este momento, estamos trabajando en el procesamiento del lenguaje natural, específicamente en el análisis de sentimientos en comentarios de vídeos de YouTube.

Al explorar los datos que podemos obtener de los comentarios en YouTube, hemos notado que muchas personas no tienen buena ortografía o utilizan abreviaciones con frecuencia. Limpiar y procesar estos textos ha sido un desafío considerable, y generar un diccionario a partir de estos datos no resulta viable, al menos desde mi perspectiva.