20/05/2025

Prueba técnica C-POCKET

Comprensión de los datos: Dado el tamaño considerable de la base de datos, se optó por utilizar una muestra representativa de 5,000 datos con el objetivo de optimizar tanto la eficiencia computacional como la precisión del análisis. La base de datos está compuesta por tres variables principales: el artículo completo, su respectivo resumen y una llave primaria que lo identifica de manera única. Tras una revisión preliminar, no se detectaron valores nulos ni registros duplicados.

Cuando se trabaja con una base de datos de gran tamaño en tareas de procesamiento de lenguaje natural (NLP), no siempre es recomendable utilizar el conjunto completo de datos desde el inicio. Es aconsejable comenzar con una muestra representativa más pequeña, lo que permite reducir los tiempos de procesamiento y facilitar la experimentación, especialmente en fases iniciales de desarrollo o prueba de modelos.

La clave primaria no presenta anomalías y se encuentra en buen estado. No obstante, las otras dos variables sí presentan inconsistencias. La columna "highlights" contiene el carácter "\n" en todas las filas, lo cual no es adecuado y debería corregirse. Por otro lado, la columna de artículos incluye símbolos como "--", que también deberían ser eliminados para garantizar la calidad de los datos. Otro aspecto relevante es la presencia recurrente de la frase "E-mail to a friend" casi al final o al final de los artículos; esta frase, junto con el contenido que le sigue, no aporta valor y resulta redundante

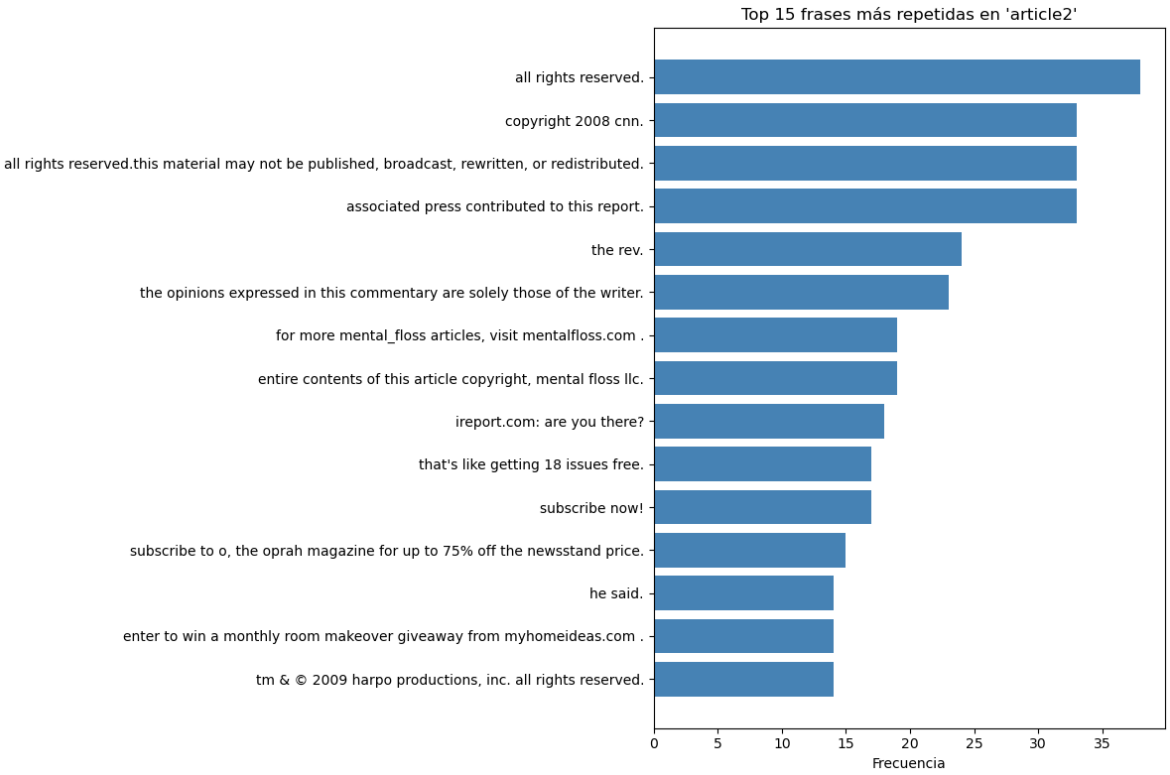
La extensión de los artículos varía considerablemente, con longitudes que oscilan entre 15,925 caracteres en el caso más extenso y 48 caracteres en el más corto. Esta amplia dispersión sugiere que la columna "highlights", que corresponde al resumen de los artículos, también presenta una variabilidad significativa en su longitud.

Limpieza y transformación:

a. Variable "highlights": Se elimino el carácter "\n", espacios innecesarios y, "NEW:"
b. Variable “id”: No presenta anomalías; no requirió ninguna limpieza o transformación.
c. Variable “article”: Se eliminó la frase "E-mail to a friend" y todo lo que le seguía hacia la derecha, ya que a continuación aparecían las contribuciones, el aviso de Copyright 2007 o los derechos reservados. Se separó en dos columnas utilizando como delimitador el primer “—”, ya que este separaba, en una sola columna, la marca de la revista, el país y la ciudad, y en la otra, el artículo sin

de la revista, el país y la ciudad. Por último, se eliminó la variable “article” y se quedó con la variable “article2”.

d. Variable “article2”: Se elimino el carácter “—“, para posteriormente revisar cuales eran las 15 frases que más se repetían.

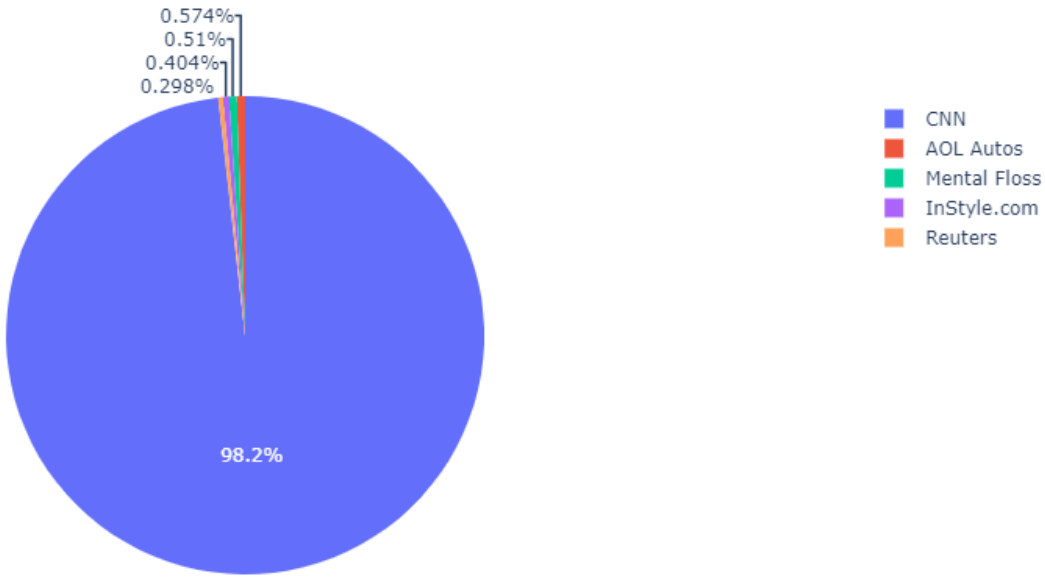


Como se puede observar algunas frases son innecesarias, y por ende se eliminan. Después de volver a correr el código, siguen apareciendo frases innecesarias y se procede a seguir eliminando, sin embargo, apareció la frase "the civil war has left more than 70,000 people dead.", que en un modelo estadístico podría ser utilizada.

e. Variable “new”: Algunas columnas en la separación por el delimitador de “-- “ se las separo mal, por lo cual fue necesario corregir. Finalmente se creó la columna “journay” donde representa las marcas de artículos.

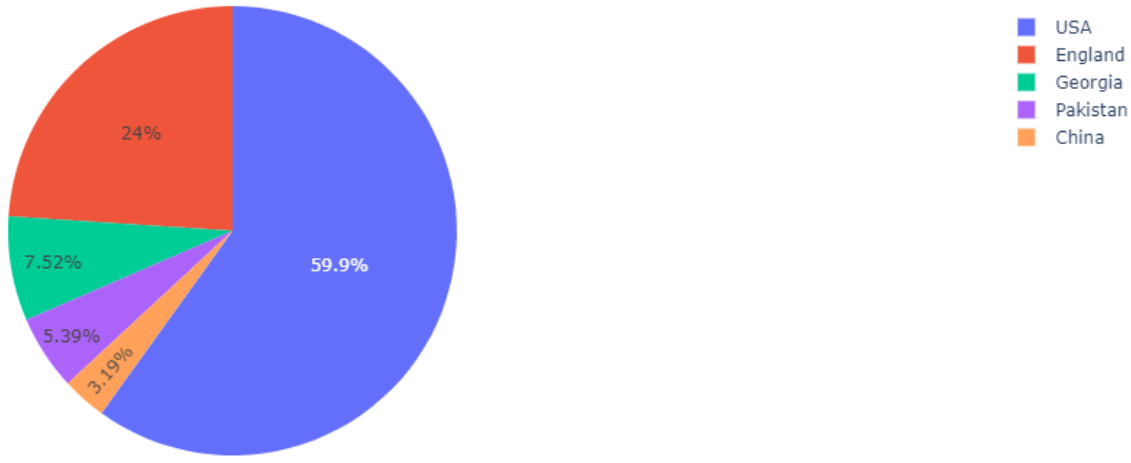
Análisis exploratorio:

5 revistas con más artículos en las páginas:



Al tratarse de una base de datos proveniente de CNN, es comprensible que la mayoría de los datos correspondan a este medio. No obstante, también se identifican otras fuentes como Reuters, aunque con una presencia significativamente menor.

5 países que generan más noticias



Dado que CNN es un medio de comunicación estadounidense, era previsible que la categoria USA predominara como el país más mencionado en sus artículos. El

segundo país con mayor presencia es el Reino Unido, que, aunque destaca frente a los demás, se encuentra considerablemente por debajo del primero.