Proyecto I: Análisis de tendencias

1. Objetivo

El objetivo de este proyecto es predecir qué palabras clave debe de tener el título de un vídeo para que triunfe, en Canadá. Cómo objetivo secundario, podríamos ver cuales son el número de palabras óptimas para el título de un vídeo, de modo a tener más visitas.

2. Extracción de datos

Estos datos los vamos a obtener de un Dataset preestablecido, de la página "Kaggle"; https://www.kaggle.com/datasets/datasnaek/youtube-new

3. Preparación de datos

i. Limpieza de datos

Realizamos un vistazo general de los datos para poder seleccionar las columnas adecuadas para este proyecto. Buscamos valores físicamente imposibles, por ejemplo, visitas negativas. Seguidamente, buscamos valores nulos de modo a eliminar las filas correspondientes. En este dataset, nos interesa la columna "videos eliminados o erróneos", en el caso de que la columna se identifique como "True", eliminaremos la fila.

A continuación, como queremos realizar un estudio de tendencias, necesitaremos filtrar y limpiar los datos de texto de símbolos raros o espacios inútiles.

ii. Tokenización

Seguidamente, tokenizamos las frases, es decir, las dividimos en palabras. Opcionalmente podemos realizar una visualización gráfica de qué palabras se repiten más. En este punto, podemos eliminar aquellas palabras que actúan como "outliers".

4. Exploración de datos

i. Gráficos Simples

En esta sección, ya tenemos los datos filtrados y procesados, podemos empezar a visualizarlos con gráficos simples, como un "Barplot" o un "scatterplot".

ii. Gráficos más complejos

En este tipo de estudios podemos realizar una exploración visual un poco más compleja, como por ejemplo, la representación "Wordcloud".

Resumen del Proyecto

Tenemos un dataset que se compone de los datos de las tendencias de youtube en Canadá (Títulos de los vídeos, visitas, likes, etc) . Nuestro objetivo va a ser ver si hay una relación entre las palabras de los títulos y las visitas. Es decir, ***¿Podemos obtener más visitas si seleccionamos adecuadamente las palabras de nuestro título?***

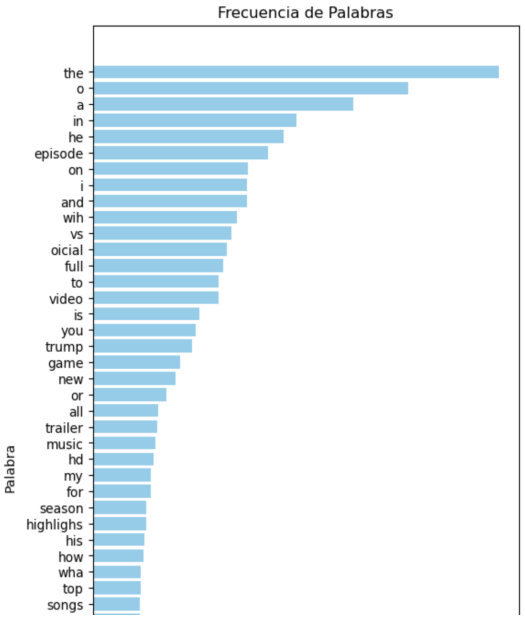
Adicionalmente, podemos preguntarnos ***si la cantidad de palabras en nuestro título influye en las visitas*** que vaya a tener el vídeo. Un título demasiado largo, puede actuar como repelente para el espectador.

Comenzamos filtrando los datos, eliminando los datos improbables o sin lógica y los que sean nulos. A continuación, tokenizamos las palabras de los títulos y realizamos una representación visual de qué palabras son las más repetidas.

Observamos algunas letras que actuarán como “Outliers”. En este estudio, no son determinantes, sin embargo podría ser adecuado quitarlas del dataset.

Llegados a este punto, contaremos la frecuencia de las palabras que se repiten en todo el dataset. Hablando, evidentemente, de las palabras de los títulos de cada vídeo.

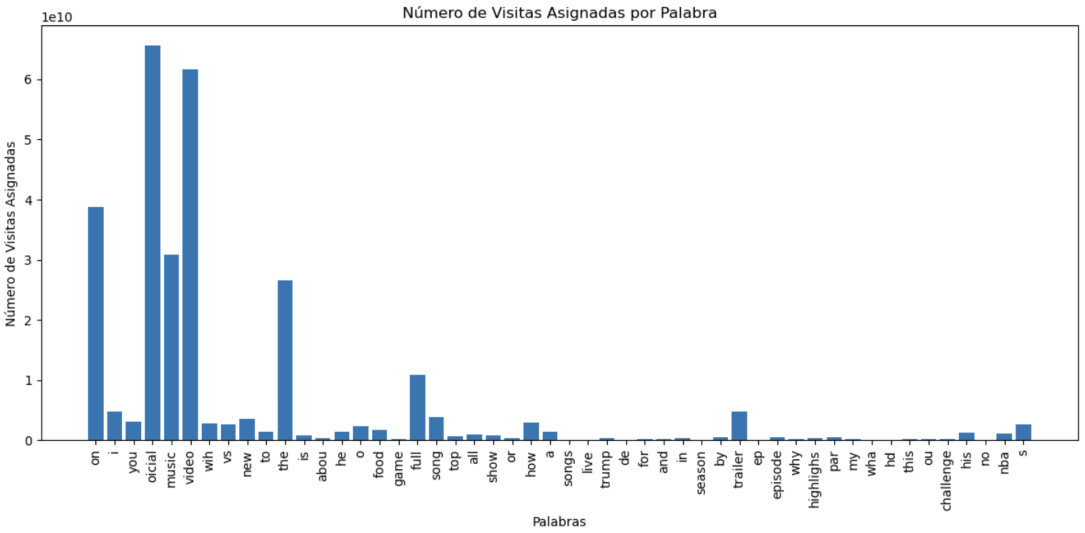
Como nuestra lista de palabras es demasiado larga sería demasiado pesada su representación. Establecemos un Umbral para reducir ese gasto de recursos. De esta manera, obtenemos las palabras que más se repiten.



Podemos identificar que las palabras, significativas, que más se repiten son: “episode”, “trailer”, “Video”, “oficial”, etc.

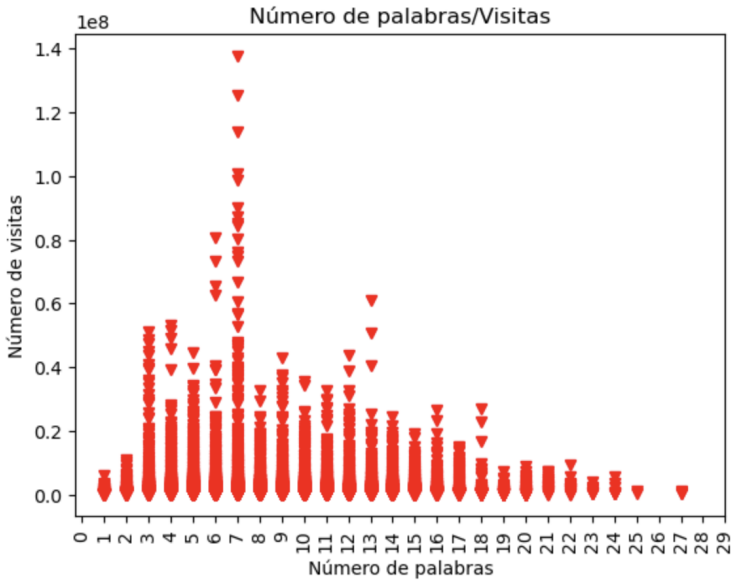
Lo siguiente que trataremos de hacer será asignar las visitas de los vídeos a las palabras que más se repiten. Es decir, hacemos una lista de palabras destacadas, filtrada con el umbral establecido anteriormente, y la vamos comparando con las palabras de cada título del dataset. En el momento en que coincidan, le asignaremos las visitas de los vídeos.

En resumen, estamos asociando las palabras más repetidas a sus visitas, así veremos si hay una relación entre las visitas y las palabras a escribir en el título.



Concluimos con que las palabras más significativas van a ser: “oficial”, “video”, “music”, “full”, “trailer” y “song”. Dicho de otro modo, en Canadá, debes poner algunas de estas palabras para aumentar las visitas.

Por último, podemos ver si hay una relación entre el número de palabras de tu vídeo y las visitas que tiene.



Observamos un número de 7 palabras.