Project I: Keyword Analysis

1. Objectives

The aim of this project is to predict the most effective keywords that should be used in a video's title to ensure success in Canada. Additionally, we will investigate the optimal length of a title to increase views.

2. Data research

This collection of data was collected on the web page “Kaggle”;
https://www.kaggle.com/datasets/datasnaek/youtube-new

3. Data preparation

i. Data cleaning

We will begin by examining the dataset in detail. Cleaning the dataset is the next stage, which involves removing extraneous characters, redundant spaces and other extraneous elements. Relevant columns for this project will be selected, and values that are physically impossible or inconsistent will be identified.

ii. Tokenization

Finally, tokenization will be performed to ensure accurate analysis. We need to segment the sentences, dividing them into individual words. If needed, we can use a graphical method to display the outcome. Any words that are not required can be removed. After that, we can examine the data using basic charts.

4. Data exploration

i. Simple chart

In this section, we possess a set of data that has been scrubbed and filtered. We can depict this data through various plots, such as "barplot" or "scatterplot".

ii. Complex chart

By using the "worldcloud" exploration, we can generate a more intricate graph that presents the same information but is much more user-friendly.

Abstract

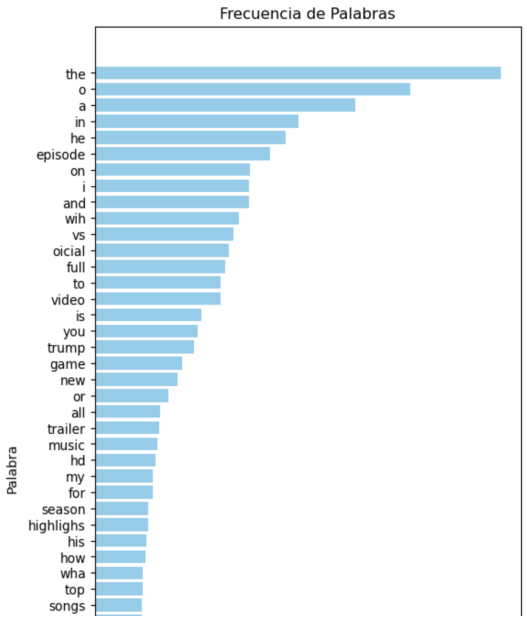
We have a dataset comprising data on videos' titles, views, likes, and more. Our primary objective is to identify a correlation between video titles and the number of views they receive.

Therefore, the initial inquiry is, **can we boost the view count by selecting appropriate words?.** Additionally, we can investigate whether the number of words in a title influences the outcome.

We will begin by eliminating irrational words, superfluous symbols or extra spaces. After tokenising the sentences and visualising the output:

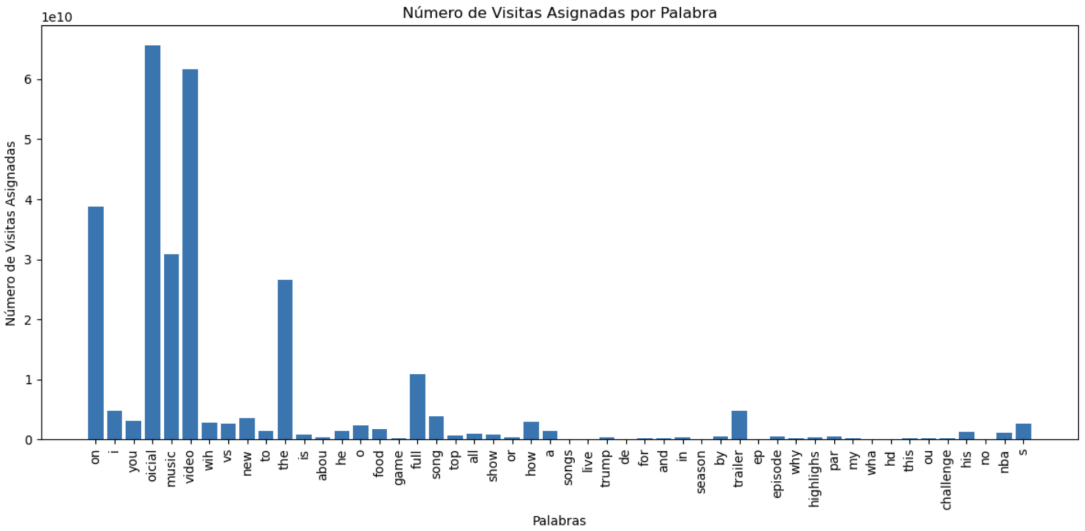
We notice some words, behaving like outliers. While irrelevant in our current scenario, it may be appropriate to remove them in other cases. Next, we proceed to count the frequency of duplicate words in the dataset.

To limit the amount of resources required, we set a minimum threshold for repeated words. The result of this



We can identify the significant words: “episode”, “trailer”, “Video”, “oficial”, etc.

The next step is to allocate views based on the most frequently used words. To achieve this, we create a filtered keyword list with a set threshold and compare this to each title in the dataset. Once this condition is met, the words will be assigned to the corresponding video views.



Consequently, the most significant keywords are: “oficial”, “video”, “music”, “full”, “trailer” y “song”.

Finally, we can observe a correlation between the number of words in the title and the number of views. Clearly, in Canada, videos with seven-word titles are likely to attract more views.