Project Report
YouTube Trending Video Analysis

CMPT 733 BIG DATA LAB 2

PROFESSOR
Jiannan Wang
Steven Bergner

TEAM MEMBERS
Jinyao (Timmy) Lu
Wanyi Su
Wei (Joann) Zhang
Yin Yu Kevani Chow
Zeyu Hu

DATE
04/13/2022

Table of Content

1. Introduction

1.1 Motivation and Background

YouTube, the pioneer video sharing platform has more than two billion users. Many companies choose YouTube as an important digital marketing platform. At the same time, more and more people choose YouTuber as their career these days. For example, Apple has spent almost 24 millions USD a year on advertisements through YouTube. A popular YouTube channel can earn up to nearly 30 millions USD annually. Therefore, a conclusion can be reached that YouTubers and business entities who would like their videos trending would care about this project.

It is meaningful to ask ourselves the question: what are the elements to be listed on the YouTube trending video list so that the videos can reach as many audiences as possible with higher numbers of views, likes, and comments? By solving this, business entities can invest wisely on specific YouTube Videos by product placement, and YouTuber can use a more systematic way for their content creation. There has been some related academic work in the trending YouTube videos. A paper published in 2021 suggests that viewers express their different levels of interests through views, likes, comments, and dislikes to various categories of videos based on videos' attributes, for example, the specific time frame they are uploaded (Khanam et al., 2021). Besides, patterns of specific video categories are examined in academia. Gajanayake and Sandanayake analyze the patterns of trending game video on YouTube with sentiment analysis, and predict trending game videos based on their attributes, such as, publish time, title lengths, views, and likes, with classification models (Gajanayake & Sandanayake, 2020).

In light of the importance of YouTube trending videos, US, UK, and Canada as three English speaking countries with a significant number of YouTube viewers are picked in this project for analysis, and a data analysis product which can show the insight of the trending YouTube videos is proposed.

1.2 The dataset

The dataset is divided into two parts: previous trending video data extracted from Kaggle (1 January, 2021 to 22nd March, 2022) and data extracted from Youtube API (23rd March, 2022 to 6th April,2022). There are 281,338 records in total including the US, Canada and Great Britain. Here are the fields in the dataset.

Fields	Data Type	Description
video_id	Integer	Id of the video
title	String	Title of the video
publishedAt	Datetime	Publish date time of the video

channelId	String	Id of the channel
channelTitle	String	Title of the channel
categoryId	Integer	Id of category
trending_date	Datetime	Date when the video is trending
tags	String	Tags used in the video
view_count	Integer	Number of views as of the data collected
likes	Integer	Number of likes as of the data collected
dislikes	Integer	Number of likes as of the data collected No data returned after 10th November, 2021
comment_count	Integer	Number of comments as of the data collected
thumbnail_link	String	URL of the video thumbnail
comments_disabled	Boolean	Enable or disable comments
ratings_disabled	Boolean	Enable or disable likes/dislikes
description	String	Video description

There are JSON files for category Id and category title which are used for mapping the corresponding information. Here are the category Ids found in the dataset.

Category Id	**Category Title**
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy

24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism

1.3 Problem Statement

In the project, three main problems are analyzed. These problems are challenging. One of the reasons is how to figure out meaningful problems analyzed that can be insightful to the practical product. Another reason is how to get the latest data that is adequate for the analysis. What's more, it is important to figure out a form of presenting the analysis results so that it is comprehensive to even a non-technical audience.

The first problem focuses on trending videos' publish date and time. The correlations of publish date and publish time against views, likes, comments counts of trending videos will be analyzed by country. The purpose of this problem is to understand how publish date and time impact the number of views, likes, and comments so as to determine the best publish date and time of videos to be listed as trending videos.

The second problem focuses on cultural differences of trending videos in three countries. Top categories and top keywords in titles and tags of trending videos in three countries are examined. The purpose of this problem is to understand what categories, keywords in title and tags are more popular to attract more viewers through more views, likes, and comments, so as to determine the essential elements of being listed as trending videos.

The third problem focuses on cultural similarities of trending videos in three countries. The purpose of this problem is to see whether there is a trend of globalization of interest brought by the YouTube algorithm.

2. Exploratory Data Analysis

EDA results for three countries' YouTube trending videos are displayed in interactive Tableau dashboards by different forms, such as, bar plots, scatterplots, and heatmaps. Quite a few videos are observed trending multiple times. Metrics, including the number of likes, views, and comments among different categories are presented in bar plots, and top numbers of categories are entertainment, music, and gaming. The number of comments are far less than those of likes and views. Correlation between likes, views, and comments is also explored. Views and likes have a high correlation of around 0.76, followed by likes and

comments with a correlation of around 0.65, while a low correlation of 0.42 is between views and comments. High correlation pairs, which are likes & views and likes & comments, are also zoomed in by scatterplots. Video trending time and publish time are also examined. Most trending videos have a time gap of 1 day to 3 days from publish to trending. For publish days of week and corresponding daytime, most trending videos are published during 10am to 10pm and peak at afternoon during weekdays, with an exception of peaking at midnight on Friday. When it comes to time of views, likes, and comments, most reactions are made by users during late nights from 2am to 5am, and among this, most reactions are made on Monday and Friday in late nights (5am and 3am respectively), Thursday at 11pm.

In the following parts, further zoom-in analysis into metrics (likes, views, comments), titles, tags, and publish time by country would be explored in detail.

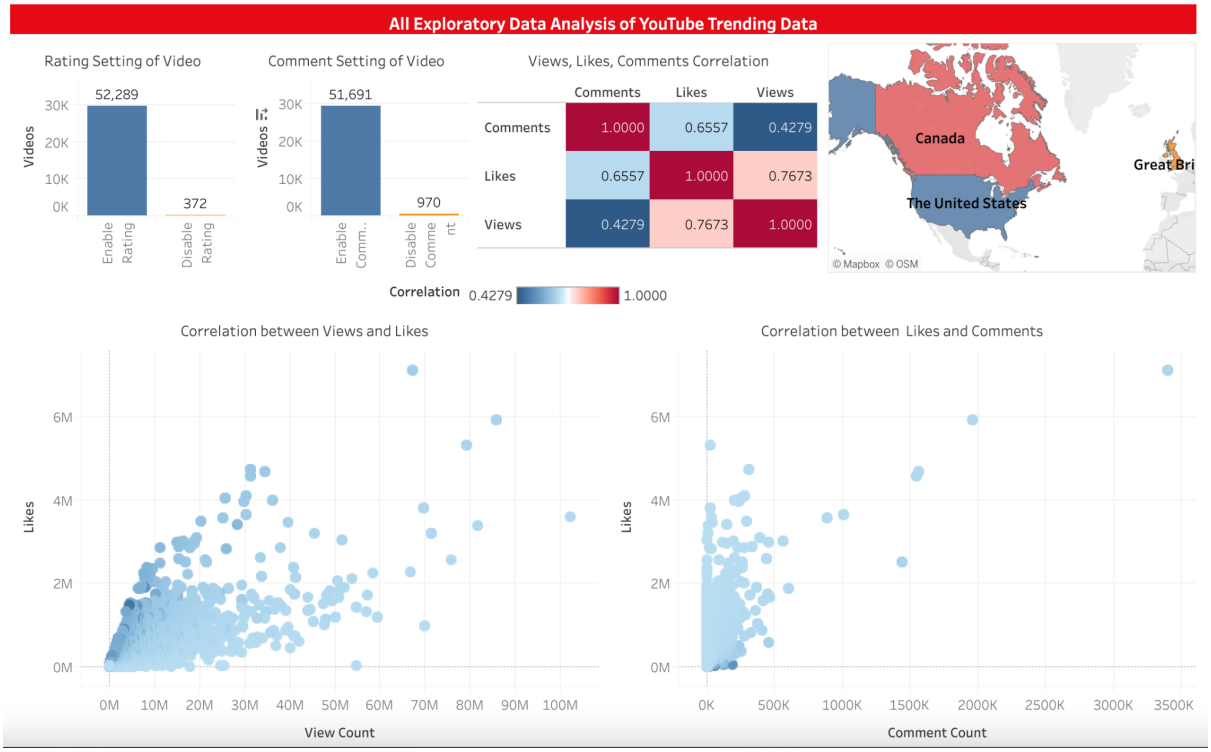

Figure 1. One of the Tableau Dashboard of EDA

3. Data Science Pipeline

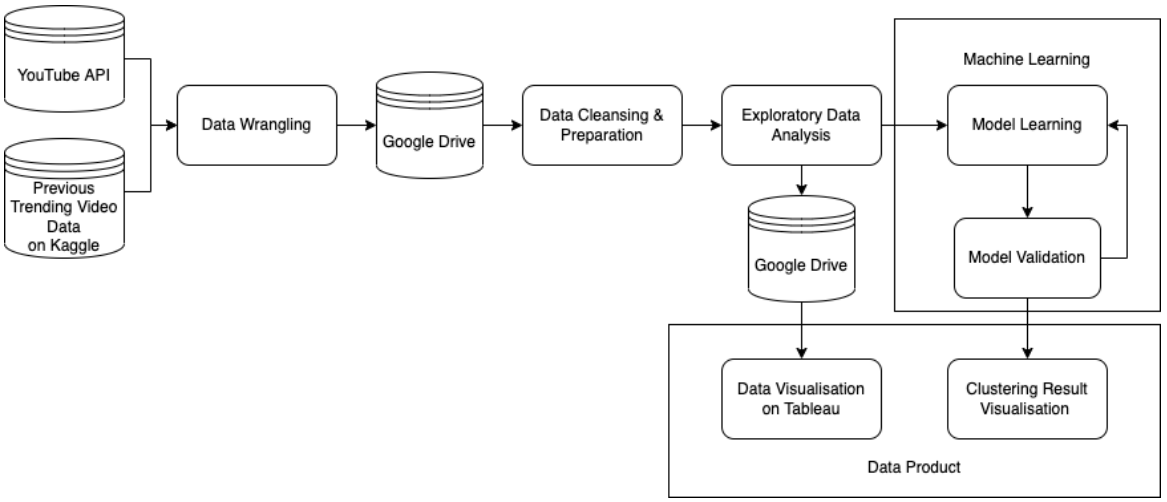

Figure 2. Data science pipeline diagram

The previous data is downloaded from Kaggle and combined with the data extracted from YouTube API running daily. The extracted data are combined and uploaded to google drive. After that, the data is downloaded by google API and continues for data cleansing and preprocessing including matching the category Id, date time adjustment according to corresponding timezone as well as extracting words from tags and titles. There are python programs for exploratory data analysis and the generated results will be uploaded to google drive through API automatically. In order to achieve interactive data visualisation dashboards with latest data, Tableau is connected with google drive dataset and published to Tableau Public as well as embedded to the data product. On the other hand, there are python programs for model learning and validation using machine learning algorithms. The results are then calculated and presented on the data product through web application. Therefore, an end-to-end data analysis product is developed with on-demand update of data using mostly automic workflow.

4. Problem Solving

4.1 Publish Date Time Analysis

After data preprocessing, the data timezone has been converted corresponding to the country. For the United States and Canada, Eastern Standard Time is adopted. Greenwich Mean Time is used for Great Britain data. Through exploratory data analysis, publish date time of the video is one of the most important elements contributing for trending. Among the three countries for analysis, the overall days between published and trendings are mostly stated on 1 days and 2 days. Therefore, in order to be selected through the YouTube algorithm for trending video, the first 2 days should be the critical moment to decide whether the video is successful. Furthermore, video count, number of likes, views and comments of each category are considered for the correlation with publish date time.

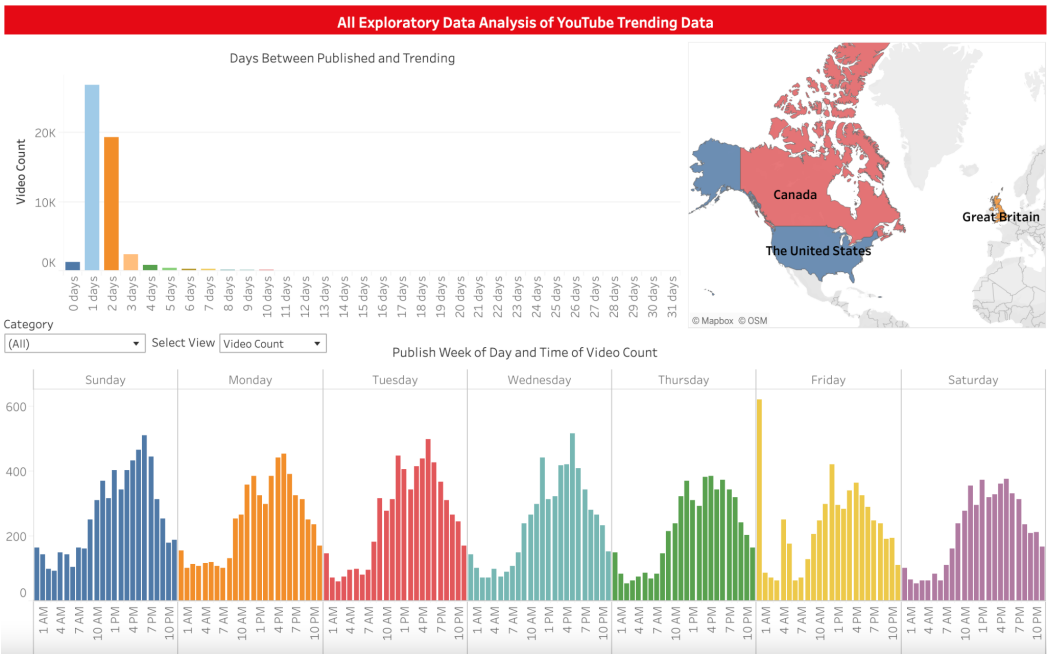

Figure 3. One of the Tableau Dashboards of publish date time EDA

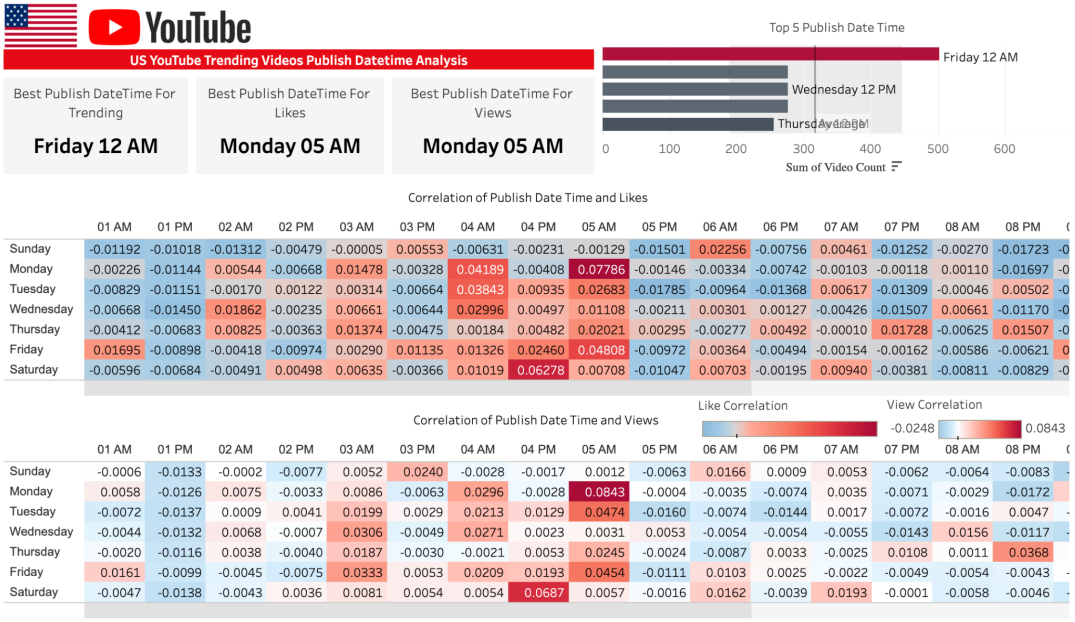

Figure 4. Tableau Dashboard of publish date time correlation analysis of US

By data visualization, there is an obvious result of publish date time correlates with video count, views and likes. Here are the results captured.

Country	Analysis Aspect	Result
US	Best Publish DateTime for Trending	Friday 12 AM
CA	Best Publish DateTime for Trending	Friday 12 AM
GB	Best Publish DateTime for Trending	Sunday 06 PM
US	Best Publish DateTime for Likes	Monday 05 AM
CA	Best Publish DateTime for Likes	Thursday 11 PM
GB	Best Publish DateTime for Likes	Friday 03 AM
US	Best Publish DateTime for Views	Monday 05 AM
CA	Best Publish DateTime for Views	Thursday 11 PM
GB	Best Publish DateTime for Views	Monday 09 AM

4.2 Titles & Tags Analysis

The analysis of the first problem shows that popular trending video categories in three countries are similar. Music, entertainment, gaming, sports, comedy, and people & blogs are the top trending video categories with the most likes, views, and comments amounts. Exploring further into titles and tags of trending videos, for country United States (US), top popular keywords appearing most in titles include “minecraft”, “fortnite”, “game”, “2021”, “trailer”, “music”, “feat” and top popular keywords appeared most in tags include “game”, “fortnite”, “challenge”, “song”, “music”, “highlight”. These keywords validate the categories with the most likes, views, and comments. Similar scenarios happen to country Canada (CA), with similar keywords as those in US, such as, “minecraft”, “game”, “fortnite”, “2021”, “highlight”, “music”, “official”, “trailer”, appear in titles, and similar keywords, such as, “fortnite”, “game”, “highlight”, “challenge”, in tags. Both countries’ viewers have a large interest in sports basketball, football, and boxing, which can be verified by keywords “NBA”, “league”, “boxing”, “UFC”, and “WWE”. Also, emoji symbols are commonly used in trending videos’ titles and tags of the three countries. Nevertheless, minor differences also exist. It implies that media like Disney and Fox are popular words among US viewers, while Twitch is more popular in CA. Apart from the same popular names appearing in both countries’ trending videos including “James”, “Paul”, and “Tom”, the US has also unique popular names like “Jake”. Besides, apart from commonly popular videos related to the keyword “Spiderman”, keyword Squid Game related videos seems to be prevalent in CA’s viewers. When it comes to the country Great Britain (GB), differences are more obvious. Apart from some popular keywords like “minecraft”, “official”, “2021”, “music”, and “trailer”, different trending keywords featuring GB’s unique geographical characteristics and

culture that are different from those of US and CA present in GB's trending videos' titles. Holiday related words like "christmas", geographical-featured words like "UK", "euro", and "England", sports football related words, such as, "Liverpool", "Chelsea", "Manchester", and "Arsenal", media related words like "BBC", and popular persons' names like "Harry", "Ronaldo", and "Solskjær" present in trending videos' titles. A similar pattern applies to GB's trending videos' tags.

The reason for the differences of the three countries' trending keywords and categories can be that GB is a traditional English speaking country, while US and CA are immigrant countries and are geographically located near to each other so their cultures are quite similar.

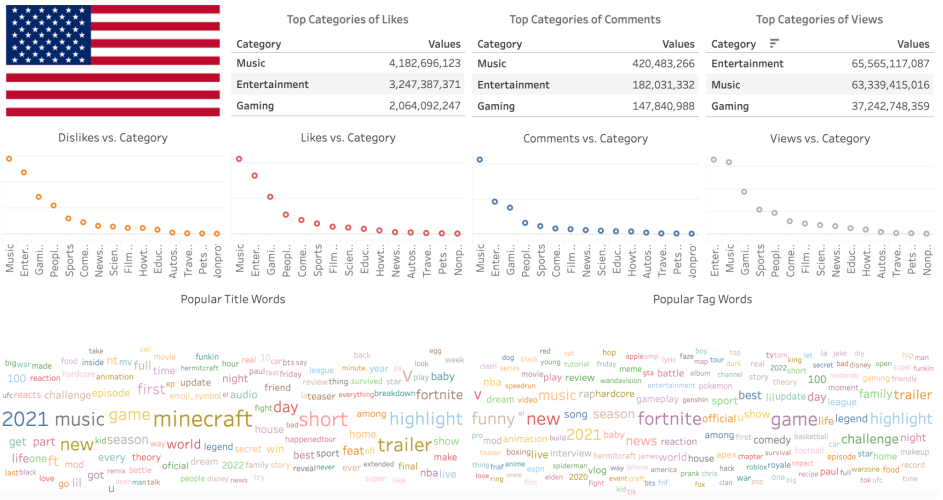

Figure 5. Tableau Dashboard of title & tags correlation analysis of US

4.3 Cultural Similarity Analysis

The analysis of the cultural similarity shows that gaming, entertainment, music, people & blogs, and sports are the top popular categories of trending videos in US, CA, and GB with numbers of trending videos higher than an average number of 552 videos. Top 10 trending channels mostly belong to gaming, sports, and entertainment categories. Similarly, popular titles and tags of the three countries' trending videos include keywords from these top categories. It can be concluded that there are a large number of fans among viewers in the three countries for popular pixel games, such as Minecraft, Hermitcraft, as well as other games like Fortnite and Pokemon. Music-related words are also important parts of titles and tags to catch people's attention, which can be verified by keywords like "music", "song", "ft", "feat", "mv". Among them, Korean pop music stands out, with representatives like "BTS". Episodes and movies seem to be a focus when people watch YouTube since "movie", "season", "trailer" occurs frequently in trending titles and tags. Besides, the prevalence of smartphones makes normal people share their life more easily and popular keywords, such as, "vlog", "home", "house". Also, people increasingly tend to share their attitudes towards news. For example, one of the hottest news topics people discuss is the war in Ukraine, which can be seen as the keyword "war" in many related trending videos' titles and tags. What's more, people in the three countries watch "official"-tagged videos frequently.

Furthermore, comparison with countries one by one is made. For geographically closely located countries US and CA, culture and viewer preference are quite similar. Top 10 channels mostly belong to sports categories, within which basketball channel NBA and fighting channel UFC are the most popular trending sport keywords. Since the two countries are “nations on wheels”, automobile channel “Donut Media” is also in top trending channels. Similarly, US and CA YouTube viewers love to watch videos with titles and tags related to the games we mentioned before. For CA and GB, of whom the former country was closely influenced by the latter in the history, cultural similarities are shown. Sports channels, such as, LaLiga Santander, Emirates FA Cup, Formula 1, and Windies Cricket, and entertainment channels, such as, United Stands, Calfrezy, Liam Thompson, are trending in both two countries. People in these two countries who share a sovereign, Her Majesty Queen Elizabeth II, are passionate about sports types different from those of the US and GB who share geographical locations in North America. They have a great passion for soccer, cricket, and F1 since 4 of 10 top trending channels fall in these types. Besides, the viewers like entertainment channels. Another interesting fact is viewers in US and GB value videos of a short length since a high frequency of keyword “short” in tags.

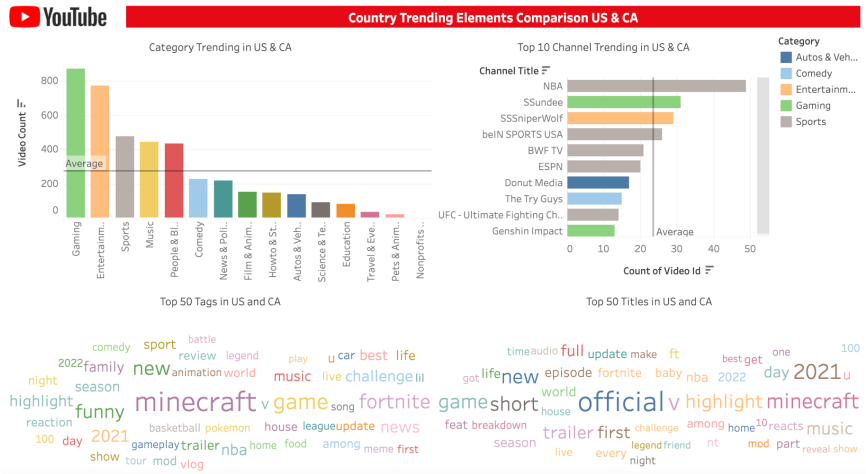

Figure 6. Tableau Dashboard of cultural similarity analysis of US & CA

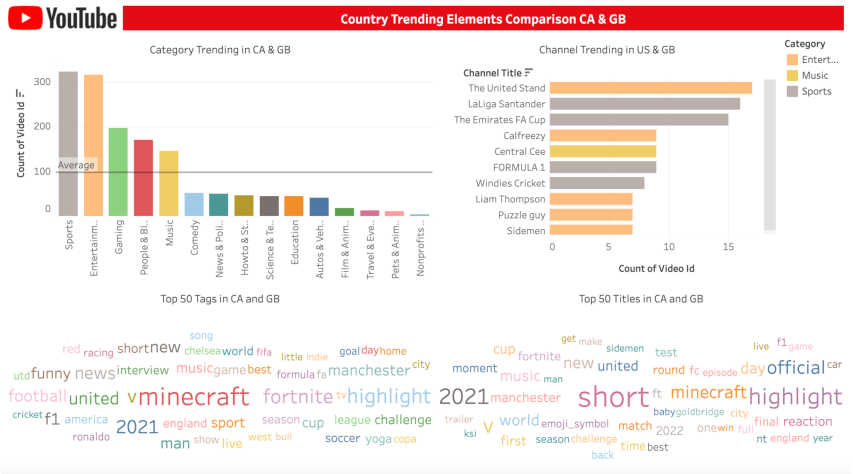

Figure 7. Tableau Dashboard of cultural similarity analysis of CA & GB

5. Methodology

5.1 Tools and Analysis method

In this project, python is the main programming language adopted for data processing and analysis. For data visualisation, Tableau is chosen for presenting the data in interactive dashboards. Here are the list of python libraries used:

Libraries for data	Libraries for API
• panda • numpy • string • emoji • functools • operator • matplotlib • seaborn • json datetime • pytz • os • wordcloud • re • sklearn.preprocessing • sklearn.cluster • tqdm • nltk	• requests • pickle • collections • apiclient.discovery • google_auth_oauthlib • googleapiclient • google.auth.transport.requests

The data is extracted from YouTube API and then uploaded to Google Drive by API through corresponding libraries and authentication.

Tags and titles are processed into words by nltk libraries and unrelated stopwords are removed. Publish date time is grouped into day of week and nearest hours. Furthermore, correlation coefficients of likes, views and comments are calculated. All the above data are exported into csv files and connected with Tableau for data visualization afterwards.

On the other hand, in order to present the findings in a symmetrical way, a web application is developed with HTML, CSS, javascript and jQuery. Tableau dashboards and results generated from machine learning models are embedded and organized.

5.2 Machine Learning

The dataset consists of a list of trending videos with the top number of views and likes. It can be less meaningful to draw insights via supervised learning such as Linear Regression. The features can be logically seen as performance for each video from multiple perspectives, and segregating the data into a set of homogeneous clusters makes more sense in generating insights. The two unsupervised learning methods used for clustering are agglomerative hierarchical clustering and k-means clustering. Both methods can be applied to organize huge amounts of data and scale up accordingly.

Generally, agglomerative hierarchical clustering measures the distance between records and form clusters with records that are close. The idea behind hierarchical agglomerative clustering is to start with each cluster comprising exactly one record and then progressively agglomerating (combining) the two nearest clusters until there is just one cluster left at the end, which consists of all the records (Shmueli et al., 2019). The major steps are: a. with each and one record form n clusters, b. merge the two closest clusters together, c. repeatedly merge the two closest clusters together. The main distance metric between two records is usually defined by the Euclidean distance. The formula is as following:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

Where i, j are indexes of the two records. The data needs to be normalized before calculating the distance since Euclidean distance is highly sensitive to scales of features. When the n clusters form after the first step, the measurement turns to distance between clusters. Specifically, the distance calculation involves records in cluster X and records in cluster Y. The methods widely used are minimum distance, maximum distance, average distance, and centroid distance. They correspond to the linkage method in the clustering algorithm: single linkage, complete linkage, average linkage, centroid linkage. It is worth mentioning another linkage called Ward's method. When more than one records are combined as clusters, the information of that cluster is represented by records within instead of only one record. There is therefore loss of information about individual record. The ward's method uses "error sum of squares" to measure the difference between individual records and the group mean.

K-means clustering is not a hierarchical method by defining the k numbers of clusters prior to the algorithm starts. The goal is to divide the sample into a predetermined number k of non-overlapping clusters so that clusters are as homogeneous as possible with respect to the measurements used (Shmueli et al., 2019). The algorithm works with several steps: a. start with k chosen initial clusters, b. select k random points as centroids, c. assign each record to its closest cluster centroid, d. recalculate the centroids of the formed clusters, e, repeat step c and d until cluster dispersion increases.

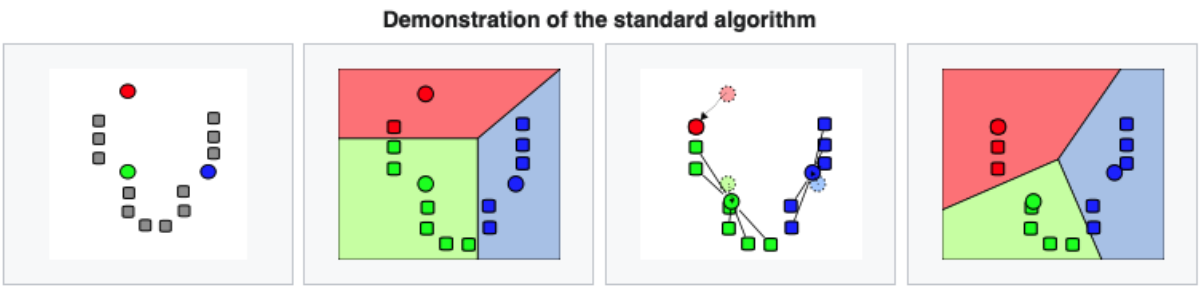

Figure 8. Demonstration of the standard algorithm | Source: Wikipedia

The YouTube trending video dataset incorporates numerical, categorical and text data. The agglomerative hierarchical clustering is applied on the numerical and categorical features. It is supposed to release insights from clusters, formed by these features and represented by their video categories. It helps narrow down the categories the product will focus on in later analysis. Additionally, the k-means cluster is applied to the text data. Title and tags of videos contain ample information and hence more promising to reveal different and valuable lessons to learn than numerical and categorical features.

5.3 Text Mining

Before modeling, it has to be preprocessed as a more digestible form so that it can be fit into the model. The first step is tokenization, to split sentences and a bunch of words into smaller pieces. In this YouTube dataset, the title and tags are combined for each video to form a long piece of text, and is broken up into separate words, lowercased, and the punctuation and special characters (symbols like %, &, $, whitespace etc.) are discarded. The second step, Lemmatization, is the process of converting a word to its base form, e.g., “caring” to “care” (Weng, 2019). This normalization process helps grasp more information about the word and similar words can be grouped together with higher accuracy. WordNetLemmatizer() from the nltk package is used in this case to keep the readable format of each word. The third step is to remove stopwords. Stopwords are those that appear repeatedly such as “the”, “we” etc. They can hardly provide useful information and the removal of them can help save computing time. Specifically in the YouTube dataset, stopwords are tailorized by adding “YouTube” and “video” that are common and less meaningful under such scenarios.

After the above process the text piece becomes a list of words for each video. Because the task is to cluster lists of words, a one-hot encoding of these words would not be helpful by creating a n-dimensional space with each word occupying one of the dimensions. The word embedding methodology, Word2Vec, is introduced to create vectors for each word with the same dimension and consequently bring context in for a list of words representing each video. Using the Word2Vec model, the vector representing each word is a shape of 1 * 50 sized array. To represent a list of words for each video, the average of words’ vector in the list is taken so that one 1 * 50 -sized vector is returned for each list. Now the data is vectorized and ready to feed into the cluster algorithms.

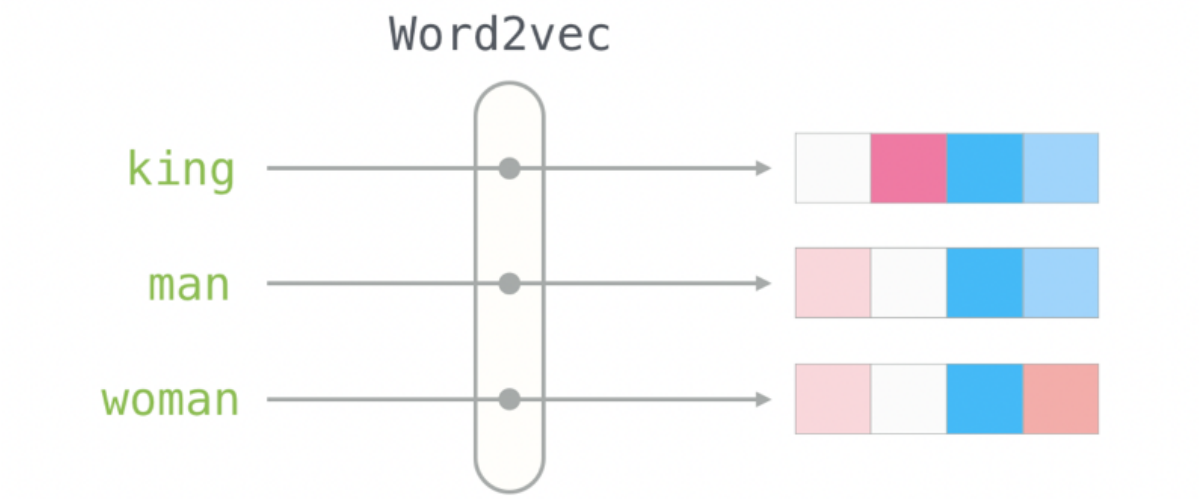

Figure 9. Demonstration of Word2vec | Source: jalammar.github.io/illustrated-word2vec/

6. Clustering Result

6.1 Clustering with numerical variables

The clustering model using numerical variables with the “ward” linkage produces the most organized output as the dendrogram shows. This dendrogram gives an idea on how the algorithm clusters the data, it shows 4 different clusters/colors of different sizes, the yellow and green clusters are very small in comparison to the red and the purple. For the sake of finding sub-clusters inside of big clusters, the cutting line is set to output 7 clusters instead of 4, by splitting the red and purple clusters. The record number in each cluster shows the cluster number 0 and 2 have the largest size, followed by cluster 4, 6 and 5.

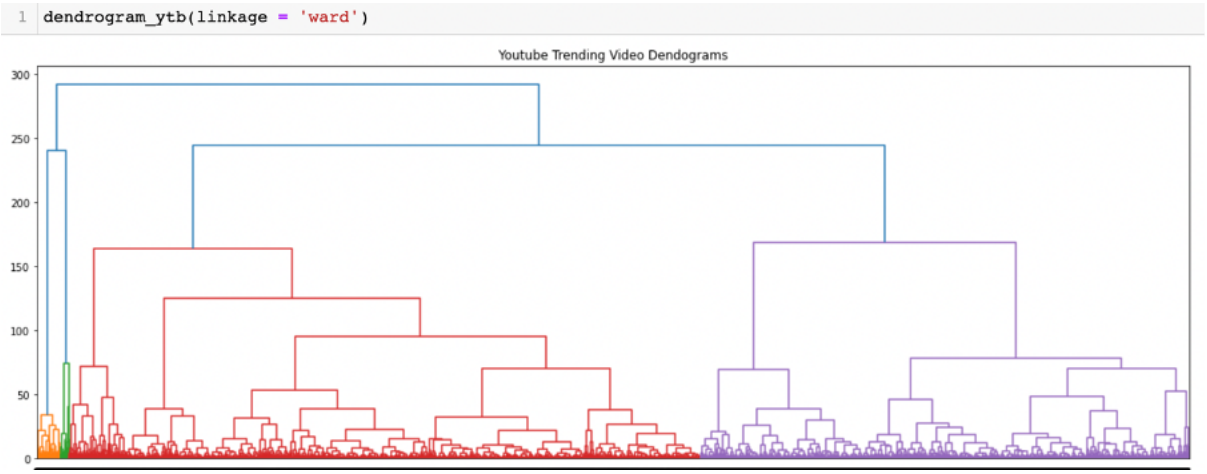

Figure 10. Dendrogram with Numerical/Categorical Features

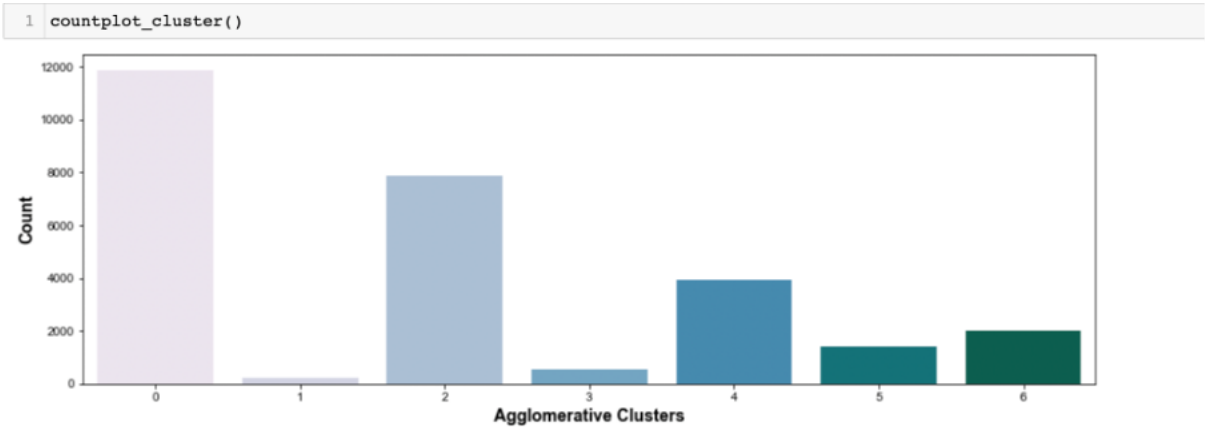

Figure 11. Result of Agglomerative Clusters

As in the dataset, only top videos with high view count and likes are listed, therefore the expected clustering result is not ideally homogeneous. This is demonstrated in the examination of numerical variables via the scatter plots. Each color represents a cluster and if homogeneously associated, there should be different colors subgroups while showing similar color within. However, the dark and light colors are mixed together in the actual scatter plots that display the relationship between numerical variables and view count. Also, there are no obvious clusters grouped together through inspection.

Figure 12. Scatter Plots of Numerical Features vs View Count

As in the dataset, only top videos with high view count and likes are listed, therefore the expected clustering result is not ideally homogeneous. This is demonstrated in the examination of numerical variables via the scatter plots. Each color represents a cluster and if homogeneously associated, there should be different colors subgroups while showing similar color within. However, the dark and light colors are mixed together in the actual scatter plots that display the relationship between numerical variables and view count. Also, there are no obvious clusters grouped together through inspection.

In the meanwhile, the analytical result of each cluster gives meaningful points from a numerical features perspective. The variable “pubToTreDays” represents the number of days it takes the video to become trending. “Likes” represent the number of likes on the day when it becomes trending. Cluster 0 trends within 1 day (median) rapidly and outperform others by winning a high number of likes (>600 million on average). Cluster 6 shares the similar behavior to trend in 2 days (median) and attracts more than 700 million likes on average. The categories in these two clusters are presented in the word cloud. The two plots logically show the same result. In both clusters, the trending videos belong to the five major categories: Entertainment, Gaming, Music, Sports, People Blogs. In the next phase of finding, the scope is based on these categories for deeper exploration via text clustering.

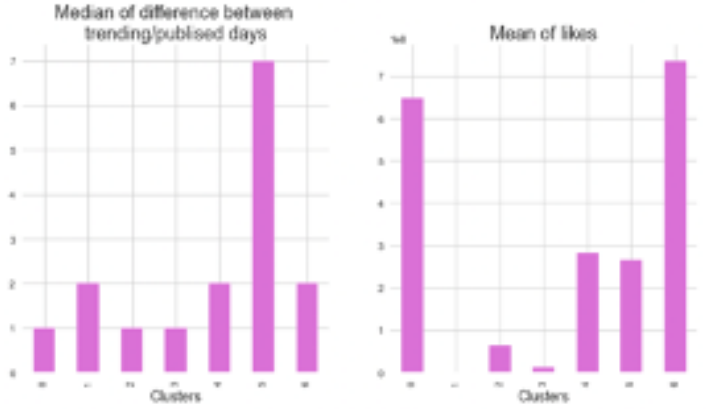

Figure 13. Median and mean of clustering result

Figure 14. Word cloud diagram of clustering result

6.2 Clustering with text

Due to that interactions among categories has created noise, the scope of the second phase finding is pinpointed to the four categories: Entertainment, Gaming, Music, Sports. Entertainment is a popular category since it is a more generic topic compared with Music or Sports. In the 4 clusters based on titles and tags, the major field revealed is on news and its close kind such as celebrity interview, official release, shows, comedy and TV seasons. Another type is various challenges, awful or funny reactions or official trailers. One special type is about football clubs and its related footballers, managers' latest news, with most of them is about Manchester United Football Club.

Figure 15. Word cloud diagram of entertainment clustering result

Gaming is a more specific topic that ranges from games info, playing to other related things. This category has been continuously increasing and the monetization potential is high. From the text clustering result, one kind of videos focus on the game news such as trailer, season releasing, update, while the major players are Minecraft, Fortnite, FIFA video game. The second kind is featured by funny moments, highlight battles, or win moments. This is a clear sub-category to focus on for gaming YouTubers who are also players with continuous materials to support. The third type of gaming videos is about review, gaming related service (twitch). The last center of attention is around techniques of game playing. The keywords are Minecraft survival, beating, challenge, dream try.

Figure 16. Word cloud diagram of gaming clustering result

The category Music is self-explanatory, including videos related to all kinds of music, songs, lyrics, instrument and teaching guide. This category is different from Gaming by incorporating subscribers from a wider range because it requires less prior background. The major type of videos is the official release of songs and albums, music clips and videos featured with another singer. Due to copyright issues, the music category relies largely on official sources and singers themselves. The other type emphasizes on the genre of music, such as rap, pop, hip hop etc. The third cluster has music topics of remix, live, and some other performances.

Figure 17. Word cloud diagram of music clustering result

Sports category includes a range of sub-categories, sport equipment, accessories, coaching and so on. This is an in-demand field and the potential earning is high. The first group is featured by basketball, with keywords of NBA, jump, offense, shot, defense. This is obviously a hot area that attracts a large number of viewers, who prefer the techniques and highlight moments of the game playing. The second group is focusing on football. There is large interaction with the entertainment category in sports as the keywords are highly related to football leagues, while in sports more specific words are brought up such as arsenal, Manchester, soccer. Another group is full of winning highlights, interviews and news.

Figure 18. Word cloud diagram of sports clustering result

7. Evaluation

Both KMeans and hierarchical clustering are unsupervised learning where there are no target variables or training and validation sets used for evaluation. Evaluating the result of clustering is unlike measuring the number of errors or calculating precision and recall in supervised learning.

The fundamental idea for clustering evaluation is on similarity or dissimilarity. Above there are several measurements mentioned on the distance between clusters such as ward, average, complete. On the other hand, it is straightforward to interpret the clustering result through statistical analysis and visualizations. Via the scatterplot there is no clear shape of segments based on the variables, except that there are 2 segments that stand out through the bar plots from the hierarchical clustering. Similar result is shown from the word clouds of these two segments by comprising the same keywords. There are other metric based evaluation methods such as Silhouette Coefficient and Dunn's Index. The Silhouette Coefficient is defined for each sample and is composed of two scores: a: The mean distance between a sample and all other points in the same cluster. b: The mean distance between a sample and all other points in the next nearest cluster (*Evaluation Metrics for Machine Learning for Data Scientists*, 2020). The score goes up while segments are well separated and dense. Dunn's Index is equal to the minimum inter-cluster distance divided by the maximum cluster size (*Evaluation Metrics for Machine Learning for Data Scientists*, 2020). Better separation also leads to a higher metrics value.

For KMeans the most popular evaluation is the elbow method. It starts with a range of integers as the candidate number of clusters (for example k from 2-20). For each k, the Within-Cluster Sum of Square (WCSS) is calculated and compared. WCSS is the sum of squared distance between each point and the centroid in a cluster (*K Means Clustering | K Means Clustering Algorithm in Machine Learning*, 2021). When plotting k against its WCSS, it looks like an elbow. Intuitively the WCSS metrics will decrease as the number of cluster

increases, but the magnitude of decreasing will be slower and slower. The following elbow plot shows at the point 4 there is a rapid drop, after which the shape moves to close parallel to the X-axis. Therefore, the number 4 (4 clusters) is selected. In most of the categories the elbow method shows similar results. Afterwards by inspecting and analyzing the keywords results in each cluster, satisfied homogeneity can be observed if returned to the cluster result section.

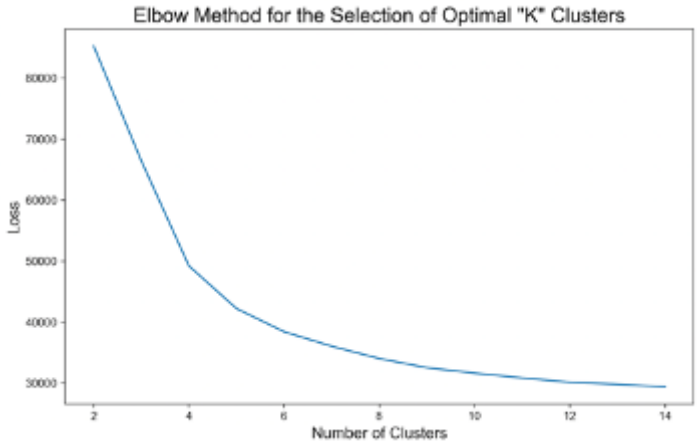

Figure 18. Elbow method for the selection of optimal “K” Clusters

8. Data Product

After analyzing various aspects of sources, YouTubers are provided with a website. From this web, users can get the following information:

- EDA results for three different countries. Users can click on the button to achieve the related Tableau dashboard result.

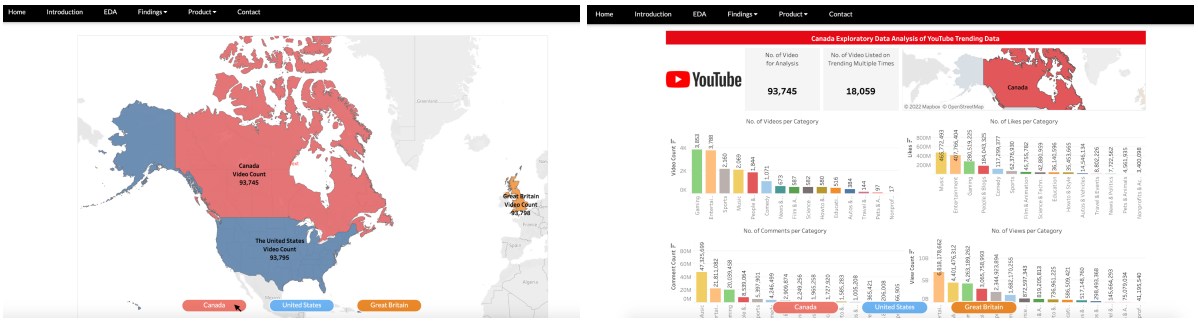

Figure 19. EDA pages of the data product

- Finding results from three aspects.

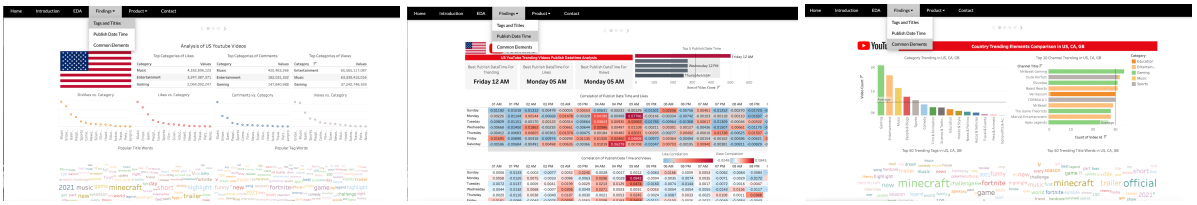

Figure 20. Findings pages of the data product

- Modeling Result and Product. YouTubers will be able to choose the category in which he is interested. This web will provide the relative text analysis and several links to the trending videos for users' reference.

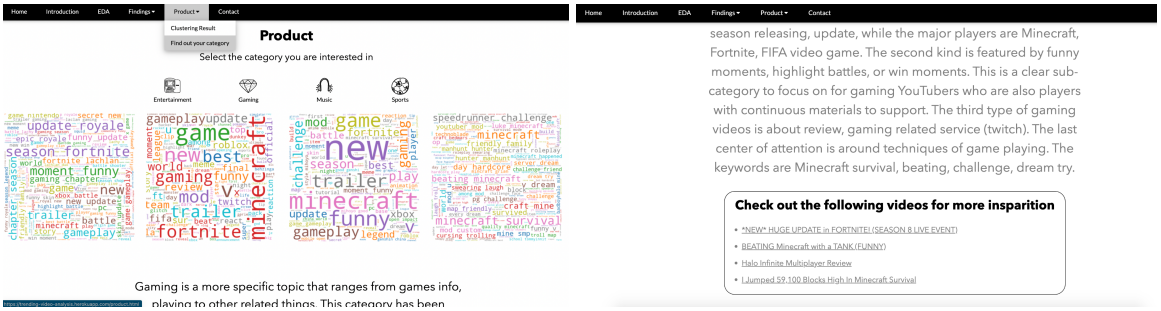

Figure 21. Product pages of the data product

The goal of this product is to help YouTubers to produce more trending videos.

9. Future Work

In the future this project will focus on techniques such as image analysis to explore the color combination and text to learn elements in trending videos. To be more specific, one can utilize the information of images' color hue, saturation, and brightness to see what kind of images are more attractive to audiences. In other words, YouTube thumbnail images are vital in helping people decide whether or not they want to watch the video, so it needs to be good. Thus, people want to find out the patterns with some computer vision technologies. Besides that, one can use image detection such as Yolov5 to find out violent or pornographic images so that people can prevent judicial or social problems.

Furthermore, some video details are worth to analyse such as duration, channel name and channel subscriber. However, in this project, due to the limitation of API calls this information is difficult to obtain as it is returned at video level and large data requests are restricted. Further investigation to improve the API calls functions are required.

10. Conclusion

10.1 Lessons Learnt

One importance before building a successful solution is to understand the business needs. If the business objective is to build an applicable model, it is pivotal to investigate multiple aspects of the problem. For a machine learning model, the necessary questions are what the model is supposed to predict, who are the key users and how they are supposed to use it practically. In this YouTube trending video modeling, the focus is put on text data and video categories so that users with specific background or interest can quickly locate the modeling result, and easily understand it via the result analysis and amazing trending videos as reference, so as to create their own content with confidence.

Furthermore, GitHub is not suitable for hosting large datasets. In order to facilitate the synchronization of large datasets, Google Drive API is explored and adopted in this project to help maintain data consistency among teammates and streamlining the workflow. Meanwhile, the data can be updated and linked to the Tableau by automical steps which achieve real time dashboards on demand.

10.2 Summary

Content creators always want their videos to reach as many people as possible. The underlying question to analyze in this project is what are the elements to be listed as YouTube trending videos in the three countries, USA, Canada and Great Britain. It may become overwhelming to solve the problem as a whole, thus it is broken down into three granular statements: appropriate publishing date and time, geographical culture difference embedded in video title and tags, comparison of common interests in the three countries. Users can explore and learn via interactive dashboards in the product. By narrowing down multiple categories with hierarchical clustering, the focuses are Entertainment, Gaming, Music and Sports. Through text mining, word embedding and KMeans clustering, the more granular segmentations are presented within each category. Users can navigate with interest and learn with their own needs. YouTube is the leading player of video sharing platforms with more than 2 million users. Overall this product aims to help businesses invest wisely while content creators can inspect internally and gain significant insight for YouTube video creation.

11. Reference

Gajanayake, G. M. H. C., & Sandanayake, T. C. (2020). Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis. *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. https://doi.org/10.1109/icter51097.2020.9325476

Khanam, S., Tanweer, S., & Khalid, S. S. (2021). Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis. *The Computer Journal*. https://doi.org/10.1093/comjnl/bxab142

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). *Data Mining for Business Analytics: Concepts, Techniques and Applications in Python* (1st ed.). Wiley.

Weng, J. (2019, August 30). *NLP Text Preprocessing: A Practical Guide and Template.* Medium; Towards Data Science. https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79

Evaluation Metrics For Machine Learning For Data Scientists. (2020, October 12). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/#:~:text=Clustering%20Performance%20Evaluation%20Metrics&text=Here%20clusters%20are%20evaluated%20based,then%20it%20has%20performed%20well.

K Means Clustering | K Means Clustering Algorithm in Machine Learning. (2021, January 20). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/#:~:text=Elbow%20Method,-In%20the%20Elbow&text=WCSS%20is%20the%20sum%20of,is%20largest%20when%20K%20%3D%201.