Julita Swiatek, Iya Chivileva

CA4022 Data at Speed and Scale Assignment 3 Final Report
Sunday 18th December 2022

Links:

Initial Analysis with Hive

PySpark Analysis, NLP, ML

Folder containing graphs, video of code running and google colab notebooks

1. Introduction

We will be analysing trending videos on Youtube over the period of 2 years, from 2020 to November 2022. We are interested in looking at trends over time and in different countries, specifically seeing if there are big differences in trending videos over the three English-speaking countries and if there is any seasonality in the videos that are trending. We are also interested in seeing how long a video stays 'trendy' – do big hits trend for a day and then are replaced by newer videos or do they stay on top for longer? Similarly, we will look at how long it takes for a video to reach the trending page – i.e. how many days it takes from the day the video was published till the moment it is trending.
We will primarily be working on Spark, be it PySpark or Hive. We chose to deploy our cluster on Spark and first carry out data cleaning and simple analysis in Hive, after which we proceeded to run Natural Language Processing and Machine Learning tasks in PySpark. We feel the choice of technology is appropriate, as we store all data in Spark, and only move between two different technologies (Hive and PySpark). We chose to use Hive mostly for simple analysis and data pre-processing as we feel it is best suited for queries such as 'find the top videos sorted by views' , 'find the top channels with most views' etc. We carried out more complex data analysis and sentiment analysis in PySpark, as we are more accustomed to working with Python and we feel it was better suited for ML tasks than Hive.

2. Dataset

Our dataset compiles data on daily trending YouTube videos. The dataset is updated daily (link here: https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset). We downloaded the newest version of the dataset, dated 12/08/2020 - 1/12/2022. The data includes entries for the USA, Great Britain, Germany, Canada, France, Russia, Brazil, South Korea, and Japan. We decided to only include datasets from English-speaking countries, so in the end our data comprises of three countries – the USA, Great Britain, and Canada.
Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, view count, likes and dislikes, video description, comment count and video category. The video categories are the same for all videos and are contained in a .json file. Our data did not have any missing values except for the 'tags' column – some of the videos did not have tags.

3. Analytics

3.1 Hive Analysis

After dropping null rows, we started our analysis with Hive. We processed simple queries like "what are the top videos / channels / categories in each country". We found that videos which were popular in one country were usually popular in all countries. Similarly, if a video was in the top 5 most watched videos, it was usually also in the top 5 of most liked and most disliked videos.

We also decided to check which channels are in the top of trending videos, but have a relatively low number of comments - the top of this list was Apple, which seems to have comments turned off completely, as it has 0 comments over all their videos.
In general, sports-related channels are very popular in all three countries, and take up large portions of the ‘top 10 channels’ leaderboards. NBA is popular in both Canada and the USA, while Formula 1 is the most-watched channel in Great Britain. When we compare particular channels, we can see that most top channels are watched all over the world, and it is usually music stars like BTS, BLACKPINK, Billie Eilish and Taylor Swift. Whenever famous musicians or bands release new music (or music videos), they immediately trend everywhere over the world, and usually stay on the trending page for a couple days. Interestingly, the longest-trending video was not a music video or even a famous youtuber - it was SpaceX’s livestream of the Starlink mission. Another interesting fact is that in the UK, videos don’t trend that long - it’s only a couple of days, usually less than 10, while the UK and Canada have multiple videos staying on the trending page for longer than 30 days.

3.2 PySpark Analysis

This is a short summary of our findings, as it is hard to summarize everything that we found over many days of research in such a short document. More comments can be found within our code, both in comments and in markdown cells in our notebook.

First, we ran simple analysis largely in line with code we executed with Hive, primarily to find out if our queries from Hive and Spark have the same results. We also did more complex analysis with Pyspark to compare the results with Hive.
It turns out that the top-liked videos are the same for each country. Below we present the summary of our findings:
The most Viral Videos in all 3 countries are :
"BTS (방탄소년단) 'Dynamite' Official MV (B-side)" ; "BTS (방탄소년단) 'Dynamite' Official MV" ;
Common Viral Videos in US and Canada :
"BLACKPINK - 'Ice Cream (with Selena Gomez)' M/V" ; "[CHOREOGRAPHY] BTS (방탄소년단) 'Dynamite' Dance Practice"
We also analysed 5 most popular categories for each country. It turned out that UK and US have exactly the same top categories. They are Sports, Entertainment,Gaming, Music, People & Blogs. Canada the same top categories except instead of Gaming it has Comedy.
We also checked if our data contains any duplicates. It turned out that there were no duplicates in any of the datasets.

When it comes to feature engineering, we started with creating new features from the ‘Published At’ column. It is a well-known fact that the popularity of a video also largely depends on the time of its publication. Videos posted on the weekends are likely to have more views than on weekdays because people have more free time. It also depends on the season, for example, workout videos will be more relevant in the spring and summer seasons. We performed feature engineering to create more attributes to train our ML model. We have created such new features as the day of the week, season or quarter, and we have also created such a new attribute as time betweeen video's publication and it trending. Based on our analysis of the relationships between seasonal data and categories, we found the following:
First of all, we can see differences in markets of the USA, GB and Canada. When looking at US data, there is a big difference between video category '10' (music) and every other category. Videos within this category trend almost twice as much as any other category. Specifically, videos related to music released on Fridays seem to trend the most both in the USA and in Canada. In the USA, the distinction is the biggest, as they have almost double the count of trending videos compared to second place, which is the gaming category, where Saturday’s videos are usually top-trending. Moreover, in Great Britain, it seems more videos trend in 3rd and 4th quarter over any other quarter.

Some categories , such as film and animation and autos & vehicles, only trend in 3rd quarter. This might be linked to movie trailers - perhaps they are most often released in 3rd quarter for the UK. Overall, more videos go trending in quarters 3 and 4 over quarters 1 and 2. This is universal across all countries, and is clearly visible - there is almost 10,000 more trending videos in quarters 3/4 than in quarters 1/2.
It is interesting that for GB specifically, it seems most videos go trending if they have been released during the week - or even early in the week (Mon, Tue, Wed). This is contrary to our assumption that if a video is released closer to the weekend, it is more likely to go viral. This seems to only be the case for Great Britain, as when running the same analysis for Canada and the United States, we found that videos released on Sunday and Friday respectively get trending more often. Interestingly, for the USA, videos published on a Saturday get viral the least out of all videos. It is surprising that there is such a big difference in the number of trending videos over these two days, even though there is only one day of a difference between them.
We also created a 'time until a video gets viral' column, which we calculated by subtracting the timestamp 'trending date' from the publication date timestamp. Our findings are that most videos will take quite a long time to get viral - most videos need from 50 to 100 to find themselves on the trending page. This is universal across all countries. The only difference we can find is that videos from Great Britain from category 17 (Sports) seem to get trending only 2 hours after they get released. This seems to be an anomaly though, and overall it is incredibly difficult for a video to get viral in under 20/30 hours. Our conclusion, therefore, is that it is impossible to get popular in one day, and many people have to consistently watch the same video over a period of a couple days for a video to get to the trending page.

3.3 Natural Language Processing

The main challenge of our project is textual data. More than half of our attributes contain textual data. Before training the model, in order not to lose any important information from the textual representation of each video, we had to resort to NLP. First of all, we extracted the semantic meaning of each video description using a pretrained Twitter sentiment model. Thus, the description of the video was turned into a binary attribute with values (negative: 0 and positive: 1). We carried out a quick analysis of the top videos grouped by category ID and sentiment. We sorted our results by average number of dislikes per video. It turned out that of the top 5 videos, 4 of them were classified as negative sentiment. This means that videos with negative sentiment are more likely to be disliked rather than liked.
Another textual feature that we were using is Tags column. First, we checked missing values in our data. We were fortunate and there were no missing or None values in any of the datasets. But while processing the Tags column we found out that there were some values that are written as '[None]'. Obviously there were not detected as missing values and it became one of our problems as we had to fix these manually.
Many recommendations in popular social networks such as Instagram or YouTube are most often based on tags. Users also often search for the content they are interested in by the tags. After studying the dataset, Google also tracks our search history and extracts meaningful information to present us with customised ads and present the most appropriate search results.
We found out that approximately the same categories top the trending page, so our key task was to predict which combination of tags could lead the video to the top. In the previous section we already prepared our dataset. We excluded unnecessary features and encoded our target attribute (tags column) into a vector space and calculated its norm. Due to the fact that under each video there is an individual combination of tags, it is impossible to encode it with standard one-hot-encoding. Instead we decided to use different NLP methods to turn our tag attribute into a numerical value. First of all we removed all stop words from tags. Secondly we tokenized them. We did analysis of tags attributes and concluded that each video has about 3-5 unique tags, that are just paraphrased. That is why we applied (TF-IDF) method to convert it to the vector format. We chose TF-IDF because it can quantify the importance or relevance of string representations in a document, which is textual tags for each video.

While preparing our data for running Machine Learning algorithms, we encountered a challenge in the form of our encoded tags column. We were not able to run models with the tags in the form of numerical vectors, as Spark's ML library only allows integer or float values as input into its models. Therefore, we were tasked with finding a way of encoding the vectors so that we do not lose too much of its value. In the end we ended up calculating the norm of each vector, as we felt this would be the best approach to not lose too much meaning of our vector. The tags feature (now in float form) will be our target variable. Now we are quite sceptical that our models will show good results.

3.4 Machine Learning

More detailed explanations of our approaches for NLP and ML can be found in the comments in our notebook.
We chose to use three different regression models for our predictions. Here are the results we got:
Linear Regression RMSE on test data 3.55
Decision Tree Regressor on test data 3.43
Gradient Boosting Tree Regressor on test data 3.11
From the results we obtained it can be seen that all the models do not fit well. It is difficult to predict the tags for videos. This is probably due to us having to encode the vector into numerical values. We also only had 7 features that we could train the model on, and prediction of tags is quite subjective as it is. There is no significant difference in the results across the models. However, the best Regression model turned out to be Gradient Boosting.
We also computed the feature importance from the Decision Tree Regressor. It turned out that the most important feature for our model is the number of likes.

4. Visualisation

We decided to visualise the connections between different channels, and which categories are frequented by the most channels, using GraphFrames and the NetworkX library. We felt that simply using a table to show our results does not allow us to fully understand the complexity of our data, so we implemented Graph Theory approaches and made a graph of the edges and vertices of our data. In the folder below, we included our three graphs (for the USA, Great Britain and Canada). The graphs can also be found in our notebook.
link to folder - CA4022 Final Assignment
These graphs represent connections between channels and different category id's.
It is interesting to see how some channels "bind" two categories together, as is the case with "Carl and Alex Fishing" which seems to publish videos both about Sports and People & Blogs (GB Plot). Similarly, the Mark Golbridge That's Football channel also corresponds to two categories - both Sports and Gaming (GB Plot). In the USA, we can see that VOX and ABC7 News Bay Area both bind the Science & Technology and News & Politics categories.

5. Responsibility statement

Julita carried out initial data cleaning and analysis in Hive. Afterwards, Iya started more complex analysis in PySpark as well as executing the NLP and ML parts. In the end, Julita performed visualisations using GraphFrames and the NetworkX library, and finished up with polishing the report. Overall, we think we both contributed equally (50%) to the project.
We used NLP to extract useful information about the videos (semantic analysis). We also utilised both Hive and PySpark for analysis of the videos. We found that there is little to no cultural difference between the US and Canada, and most videos, if trending in one country, were also trending in the other. Great Britain, however, has a different market and some videos very popular in the US or Canada were not as popular in GB. We did not carry out analysis based on language (as most videos were in English), and we also did not carry out analysis based on particular youtubers - as we felt it

would be more meaningful to look at all the videos from different categories, rather than focusing on specific people.