CSE6242 Team 134

John Herrmann: jherrmann31@gatech.edu
Qiang Wen: qwen36@gatech.edu
Shivam Goel: sgoel62@gatech.edu
Theodoros Spathopoulos: tspathopoulos3@gatech.edu
Venkatesh Vijaykumar: vvijaykumar3@gatech.edu
Vicente Santacoloma Blanco: vblanco3@gatech.edu

YouTube Trending Video Analysis

Introduction

YouTube is currently the most popular media platform on the web and can be accessed from almost anywhere in the world. Considering that more than 1bn unique users visit YouTube each month and 72 hours of video are being uploaded every minute, YouTube is one of the best places for advertisement and brand engagement. However, the increasing number of video uploads makes it more competitive to get the attention of a user. The success of a video is measured by the number of views and likes it has attracted and it takes a varying amount of time for a video to become popular. To understand modern YouTube viewing behaviour and inform advertisement and marketing strategies, we have to first understand the factors that contribute to a video's popularity growth.

Algorithmic approaches and methodology

We have analysed the publicly available data sets on trending YouTube video metadata to understand the key feature variables in different regions. We combined a range of machine learning techniques for clustering, sentiment analysis and relationship network exploration to generate useful data insights. We also performed a prediction analysis using the data we collected to predict future view count. The visualisation of our analysis was based on Python and Tableau.
The most important factor that differentiates our analysis from previous analyses is that we are focusing our efforts on multiple regions instead of analyzing the YouTube growth statistics in general. We have also included features, properties, and metadata of videos, which have been unexplored in previous analyses. The sentiment analysis for the titles and video descriptions helps to identify videos with negative or positive sentiment, and the innovative word cloud visualization gives a hint about the title segments which can contribute to a video becoming trending in a specific region.

Data collection

We developed a Python script to get the most popular videos for all regions. The data collection algorithm involves the following steps: First, we fetched metadata for most popular videos for all regions from YouTube API v3. Second, we parsed it. Finally, we saved it in CSV files; one per region. The initial data set we collected includes over 1.8M entries and 25 columns (ca 3GB).

Data processing

We developed three Python scripts to process the data. The data processing algorithm involves the following steps: First, it merged all the CSV files we obtained from the data collection stage into a single CSV file. Second, it removed duplicate videos. Finally, it appended the category name related to the category id, and the continent in which the region is located. The data set generated after removing videos with the same id and region include approximately 60 thousand entries.

Metadata analysis & general observations

We have used Tableau to create an interactive visualisation which demonstrates the metadata analysis of the collected videos (Fig. 1). We present details of the view count, the average duration of videos in each category for every region, as well as how long it took for a video to become viral in a region. Finally, we provide insights of the duration and the number of videos that include the word *live* in their title.

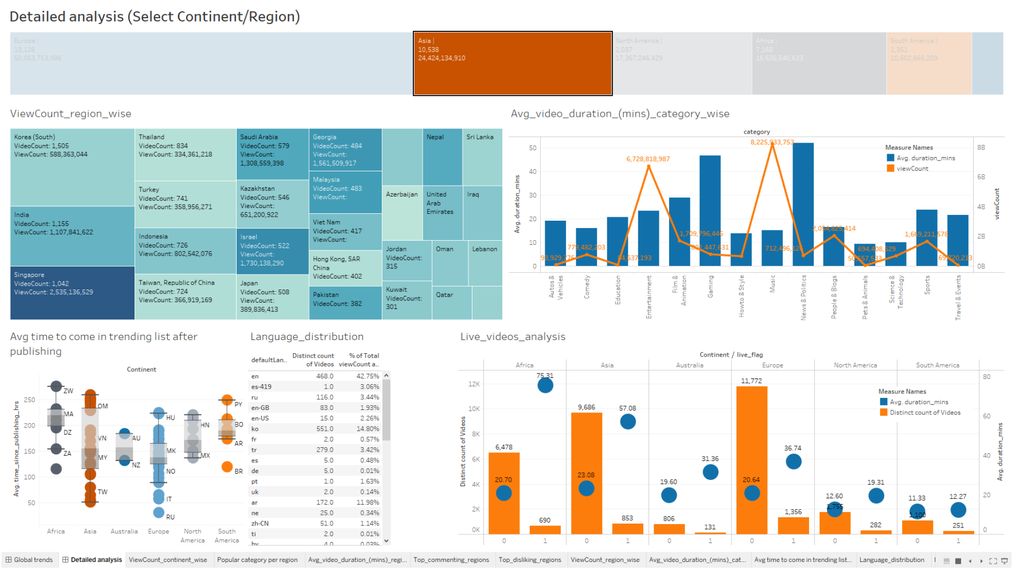

Figure 1: Screenshot of our interactive visualisation analysing the metadata of the collected videos. Top bar allows selection by continent, center left box allows selection by region. After selection of region, the center right graph shows the average video duration by video category, the bottom left graph depicts a box-plot of time between video publication and first appearance on trending, the bottom left table shows the video language distribution for the region, the right bottom graph shows the number of live vs. prerecorded videos for all regions on the selected continent.

Clustering analysis

k-means clustering in Python was used to cluster the data based on the number of views, likes, dislikes and comments, and the lengths of their title and descriptions. Clustering was performed to provide an initial insight into the structure of the data. We chose three clusters based on the elbow method, and used a scatter matrix to represent the results rather than use multiple plots or multi dimensional plots that are not very intuitive (Fig. 2). The algorithm itself is quick to execute, and is useful with even unlabeled data, and the plots help visualize feature relationships as well as data distribution. The scatter matrix below presents the relationships between the features used. The main diagonal represents the distribution of data for these features. Each cell presents the relation between the corresponding feature on the X-axis and Y-axis for that cell.

Figure 2: Scatter matrix of clustering results. Colors depict clusters. viewCount, likeCount, dislikeCount, commentCount depict numbers of views, likes, dislikes, and comments for a video. tlen and dlen depict lengths of title and description strings, respectively.

Sentiment analysis

Word cloud for video titles

Word cloud is trying to analyse the most frequent words in each trending video's title, it summarizes the frequency of each word for each region and proportionally mirrors to its map. Word cloud helps to visually understand which makes an attractive video in terms of the title, it also gives us some hint for similar trending videos in different regions.

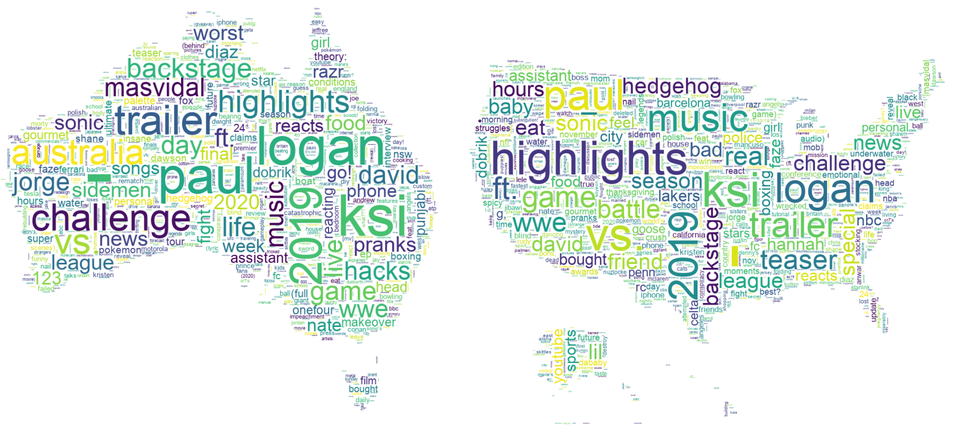

Figure 3: Word clouds for Australia (left) and United States of America (right) share the same popular words 'paul logan', 'trailer', 'ksi', '2019', 'highlights', 'challenge' etc.

Sentiment analysis for video description

A sentiment analysis was done on each video's description. The resulting compound score is a metric that calculates the sum of all lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive). The average compound value for a whole region was averaged based on the number of views for each video. The results show strong differences between regions, while viewers in Europe, South Asia and Africa mainly watched positive videos, viewers in Latin America preferred watching videos with negative sentiment.

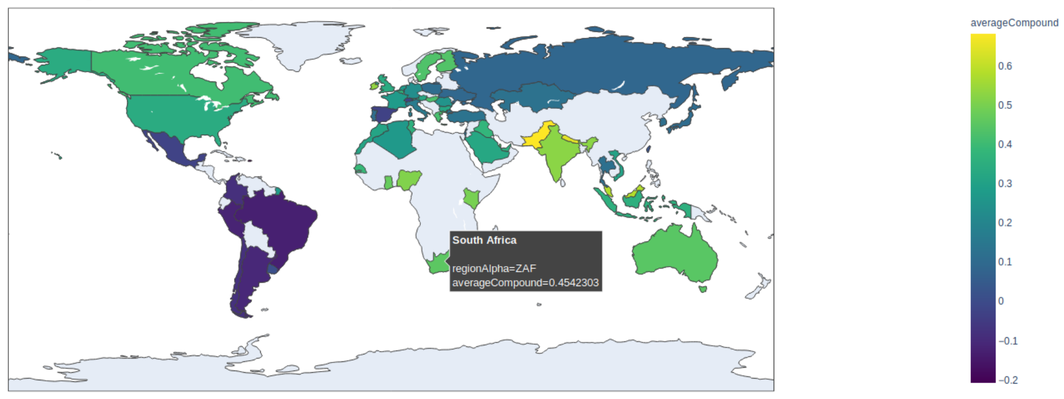

Figure 4: Map showing results of the sentiment analysis of the video descriptions for each region. Yellow indicates highly positive average sentiment, while dark blue indicates highly negative sentiment. Regions in pale blue were not included in the data set.

Network exploration

A Pearson's correlation matrix was used to investigate correlations of videos trending in different regions. The results indicate a clear correlation of trending videos between some regions, while others do not share any trending videos. Clusters of regions with the same language share many trending videos. Few major players including Russia (RU), Algeria (DZ), Canada (CA), Singapore (SG), and Argentina (AR) uploaded many videos which also trended in other regions. Videos uploaded by major players mainly trended in regions speaking the same language.

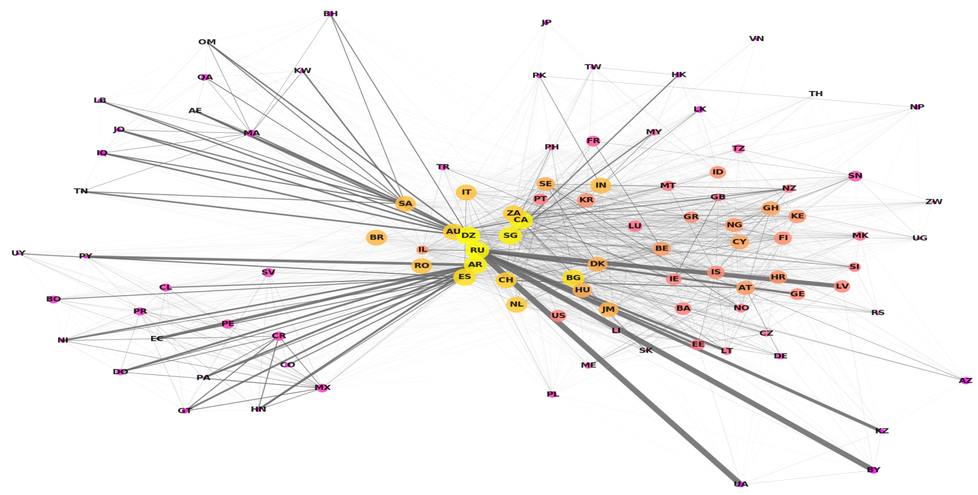

Figure 5: Network graph showing connections between regions based on trending videos. Circle size and color indicate the number of outgoing edges from a region (videos uploaded from this country which trended in other regions as well), width of edges indicate the number of trending videos shared by the two connected nodes.

Predictive analysis

We used our extensive data sets to build a neural network-based machine learning algorithm, which predicts the video viewCount given properties such as current likes and dislikes. The model was trained and hyperparameters were selected based on the test accuracy. The accuracy of the model on the test set was calculated using its R^2 score which ultimately came up to 86.75%. The neural network was chosen since it estimates highly non-linear functions well. The approach stands out from categorizing views into buckets and approaching the issue as a classification problem by approximating the actual count of views. The model, while slightly slow to train, responds well at the query and returns decent accuracy in comparison to vanilla regression models or kernel-based methods.

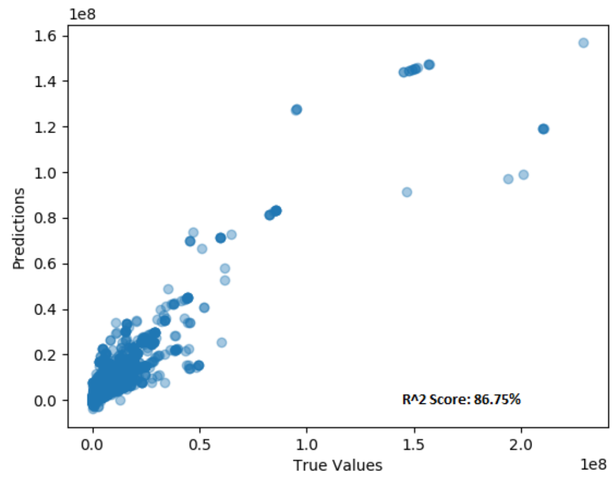

Figure 6: Scatter plot used to depict closeness of predicted results to true values of video view counts

Conclusions

- Videos from the *Entertainment* category dominated trending list in almost all regions, with *Music* coming second.
- The duration of a video on the trending list differed greatly among regions, with videos from Russia becoming viral the soonest.
- Innately numerical features provide better indications to the nature of the data, as opposed to derived numerical features especially from string based features.
- The title and description lengths do not highly contribute to determine the statistics of a video, as evidenced in the different clusters of the scatter matrix.
- The average sentiment was positive in most tested regions, except for Latin America, which showed negative average sentiment
- Regions that share the same official language tend to share a similar trending video list. A similar phenomenon appears frequently in regions with different official language but close geographic proximity
- Few regions dominate the videos on trending lists in different regions
- View count is strongly correlated with likes and dislikes. Splitting our data on a 70-30 (train-test) ratio we were able to predict view count with 86.75% accuracy