Determining North America’s Top Trending Videos on YouTube

Extract:

The data used in this project were downloaded from Kaggle.com. These datasets are a daily record of the top trending YouTube videos. They were in a CSV file format. YouTube uses a combination of factors to determine the top trending videos within a year. For this project we examined the data from the United States, Canada, and Mexico (North America). The trending dates ranged from November 14, 2017 through June 14, 2018. There were approximately 200 videos listed per date. Many videos were repeated over numerous days, as videos often trend for many days in a row. The data sets each contain a video ID, a trending date, a title, a channel title, a numeric category, a publish time, the number of views, the number of likes, the number of dislikes, number of comments, a link to a thumbnail, information about whether comments or ratings were disabled, information about whether the video contained an error, and a description.

Transform:

Data files from the United States, Canada, and Mexico were initially uploaded into a Pandas data frame. A country code variable was added to each file and coded for the country it represents (US-United States, MX-Mexico, CA- Canada). The data files were appended to each other to form one large data file. The data were then sorted by number of views.

This data was relatively easy to work with. The only issue we found while opening the data using Pandas was the Mexico data set had a formatting error. All of the files appeared to be in the same format but only this data set created an error. We were able to fix the error through defining the engine (engine='python'). The data file was also rather large and had to be zipped to load it to the GitHub repository. We found that Pandas has a command for unzipping the file and then reading it back into Pandas (NAYT\_df.to\_csv("NAYT.csv.gz", sep='\t', compression="gzip") prior to uploading it into the database.

Load:

We used a Postgres SQL database for this project. We used only one table that contained all of our data the index that was relabeled “ID” was used as the primary key. Once the database shell was created data were loaded into the SQL data base. We had some difficulty with variable names when uploading the data into our database. Some of the data was too large so we had to change the variable limits within the database. Once all of our variable names and formats were corrected the data was uploaded in the database.

We then ran a query on the database to find the top 10 videos with the largest number of views within our data table. Once we did this we realized that many videos were repeating over several days. We then ran the query using “group by title” command. We found the top ten videos with the highest number of views in North America. We used Matoplotlib to create a bar graph showing the top ten videos and number of views.