Youtube/Twitter Trending Analysis

EECS 4415 Final Project Proposal

Andres Rojas
Lassonde
York University
Toronto Ontario Canada
andres15 @my.yorku.ca

Juyoung Kim
Lassonde
York University
Toronto Ontario Canada
jyk3216@my.yorku.ca

Adrian Winkler
Lassonde
York University
Toronto Ontario Canada
adrianw@my.yorku.ca

Richmond Truong
Lassonde
York University
Toronto Ontario Canada
richt@my.yorku.ca

ABSTRACT

Youtube is an online media distribution website that specializes in videos. Youtube has a section called “Trending”, which has the most popular videos for a user’s region. Our data set is a collection of trending videos for the region Canada during the year of 2018, this will be used for batch processing. Our data set has a variety of features such as title, thumbnail, time, and others. We will be performing batch analysis on these characteristics to figure out the most important factors as well as how to optimize a video’s presentation to maximize views. In addition to the data set, we will be collecting our own streaming data on videos that are currently trending on youtube. Using the twitter API we will see how popular youtube tags are on the twitter platform. Analyzing these factors will provide us on insight into general social media trends and tags.

KEYWORDS

Youtube, Trending, Viral Media, Twitter Hashtags

ACM Reference format:

Andres Rojas, Juyoung Kim, Richmond Truong and Adrian Winkler. 2018. Youtube/Twitter Trending analysis:EECS 4415 Final Project.

1 Domain Description and Motivation

We are creating an analytics architecture that is composed of a variety of components, composing of streaming data, batch processing, and image analytics. The overall architecture of the Analytics Platform can be see in Figure 1. We will build a web scraper that collects currently trending videos as streaming data. Our analytics platform extracts the tags for these videos for further use. The extracted data is also saved to external file for batch processing later. Using the youtube trending tags we scour tweets which have the same tags and compute simple popularity scores. These two methods of streaming data are processed and saved to a external file for more in-depth analysis.

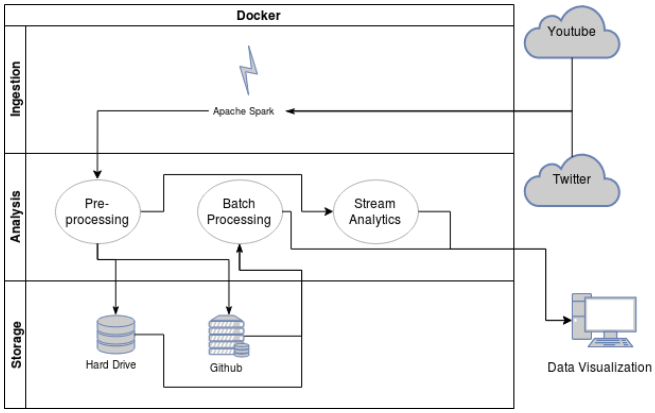

Figure 1: Analytics Platform Architecture

1.1 Data Collection and Ingestion

Due to the volatile nature of “trending” media, Apache Spark will be used to collect streaming data from youtube’s trending page as well as Twitter’s recent tweets. This data will be analyzed at the time of collection and then saved as csv files for later batch analysis. The historical batch data will be stripped of redundant information such as common tags and stop words, then pre processed into text files for more faster analysis.

1.2 Storage

All data we ingest will be saved on our personal machines as well as the PRIMS servers. A private Github repository has been set up as a backup for any code and datasets we generate.

1.3 Serving and Visualization

Once the batch data has been processed and analyzed, we will plot out simple video data (i.e. most used tags, tags that received the most views, etc). Here we will demonstrate the effect certain criteria have on a video's popularity.

Results from the streaming data will be displayed in real time, plotted to a graph showing which trending videos had the most response from twitter.

1.4 Limitations

Since this data is being saved onto non-dedicated hard drives, there is a limit to how much data we can hold. Although Github has no hard limit to available space the PRISM lab servers do. There's also a very limited timeframe to complete this project. This will limit the depth of our analysis and prevent us from implementing technologies not covered in class.

Another limiting factor is youtube's API. Although it has many implemented features, none of them seem to be relevant to our needs. Because of this, a scrapper will be implemented to retrieve trending data from youtube.

2 Architecture of Proposed Solutions

Youtube and Twitter have ways to categorize videos using tags. Youtube defines them as tags while Twitter defines them as hashtags. Our project will find some tags that are in the trending of Youtube and try to find associated tags in the tweets of Twitter. This will be done by using both the Twitter and Youtube APIs to access two stream of data and do analysis. Youtube data would be updated daily due to trending videos being updated daily while the twitter stream is in real time. This will allow us to see how often a video in Youtube trending and its associated tags are tweeted in Twitter and also see how common the interests are between users in both platforms. It will also tell us if a commonly tweeted youtube video will increase its chances of remaining on Youtube trending. A major issue that may occur during this analysis is the possibility that different hashtags are used in both platforms for the same things e.g. "soccer" and "football" vs "football" and "american-football".

2.1 Thumbnail analysis

Using the thumbnails we can see how popular a certain colour is when compared to the others. This process is called RGB analysis, it finds the most common used color in an image, and finding the most common color across all the trending images gets us the most appealing colors. Using this information and an emotion chart we can speculate what emotions are evoked and find the best emotion to evoke.

In addition to doing RGB analysis, we also have the potential to perform machine learning object detection on the thumbnails to determine the most commonly used objects in the thumbnails.

3 System Evaluation and Data Analysis

We will evaluate our system and architecture by verifying that we get the correct data. Results that we hope to obtain are showings of a direct link between the tags on Youtube and the tags on Twitter. We also hope to obtain more insight on what causes a video to go trending on youtube. The types of data analysis we will perform are Jaccard index, TFIDF, and object analysis. Using Jaccard index we will be able to see how similar the tags are on different Youtube videos, also how how many overlapping tags on Youtube and Twitter. TFIDF will be performed on the descriptions on videos. Object analysis will be done on the thumbnails of the youtube videos to look for trends in the thumbnails. We can scale our solution by gathering more data and we are doing that to gather more data for our data set. We have set up a scrapper with the youtube api and every 15 minutes when the trendings page updates, we grab the latest trending and put it into our data. Another way to scale our solution would be to include tags from other social platforms.

ACKNOWLEDGMENTS

Except for the stream data we collect from twitter and youtube, all other datasets have been collected and curated by the open source data community at Kaggle. Although we will be performing analysis on this data, we do not own any of it. All data used is either collected from Youtube, the Twitter API, or the Kaggle Dataset.