Project Proposal

Arjun Mahadkar
B00976832
amahadk1@binghamton.edu

Deepang Raval*
B00924269
draval1@binghamton.edu

Rahul Verma
B00892091
rverma4@binghamton.edu

Sudeep Rawat
B00852066
srawat1@binghamton.edu

Yuraj Vartak
B00866245
yvartak1@binghamton.edu

1 DATA SOURCE DESCRIPTION

YouTube[1] is an online video sharing and social media platform that allows users to present their views on a specific video through its comments section. It is responsible in generating vast amount of textual data in various domains like product reviews, product recommendation, event opinions etc. While there are comments that are related to the video, there are comments that are toxic, spam, self-promotion or unrelated to the context of the video collecting which can help creating filters to hide or discard them. Since there are various video categories, this Youtube comments dataset can also be used to train machine learning/deep learning models for domain specific tasks.

2 DATA EXTRACTION

We can collect the YouTube comments using an API developed by Google developers.

2.1 Resources

YouTube Data API[2] will be used to gather the required information. The API could be used by generating an API key through a Google Cloud Platform (GCP)[3] project. Each project could have a single key hence multiple projects could be created to fetch more data on a daily basis.

2.2 Methodology

(1) Choose the YouTube channel whose videos we will use to fetch the comments. New channels could be added dynamically
(2) The API supports list method for search that could help in creating a list of videoId for each channel
(3) The result json generated from step 2 also provides snippet property that helps in extracting a video's title and description
(4) Further the videoIds produced in step 2 could be iterated for list method of CommentsThreads resource to grab all the top level comments for a particular videoId
(5) The result json produced in step 4 also has replies property through which we can access all the replies to a particular comment
(6) We can use MongoDB to store the data

[1]https://youtube.com
[2]https://developers.google.com/youtube/v3
[3]https://cloud.google.com/gcp

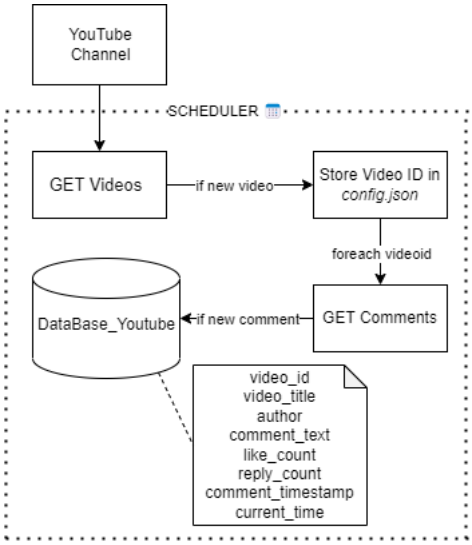

Figure 1: Methodology

Figure 1 shows the architecture of the YouTube data collection system.

3 DATA ANALYSIS

(1) Visualizing the frequency of comments vs the lifespan of video till the specified date
(2) Visualization of sentiment of people towards videos of a particular channel
(3) Count of toxic comments for each video
(4) Plot of relatable vs unrelatable comments for each video that has description
(5) Count of spam comments for each video

4 DATA COLLECTION ESTIMATES

The Youtube Data API works on a point system where each project is allocated 10,000 points per day. 1 read operation that retrieves a list of resources costs 1 unit. We can allocate 20 points to fetch the videoIds. Remaining 9,980 points can be used to fetch an average 50 comments per request resulting to 9,980 x 50 = 499,000 comments per day per project.

* https://scholar.google.com/citations?user=MfAFu6QAAAAJ.