Advanced Database Topics
COMP – 8157

YouTube Data Analysis

Biweekly Report

Submitted to: Dr. Pooya Moradian Zadeh

Submitted by: Team Lannisters

Sanyam Sareen (SID-105192200)

Karan Singla (SID-105172747)

Amrita Dhir (SID-105173647)

Lavish Handa (SID-110006198)

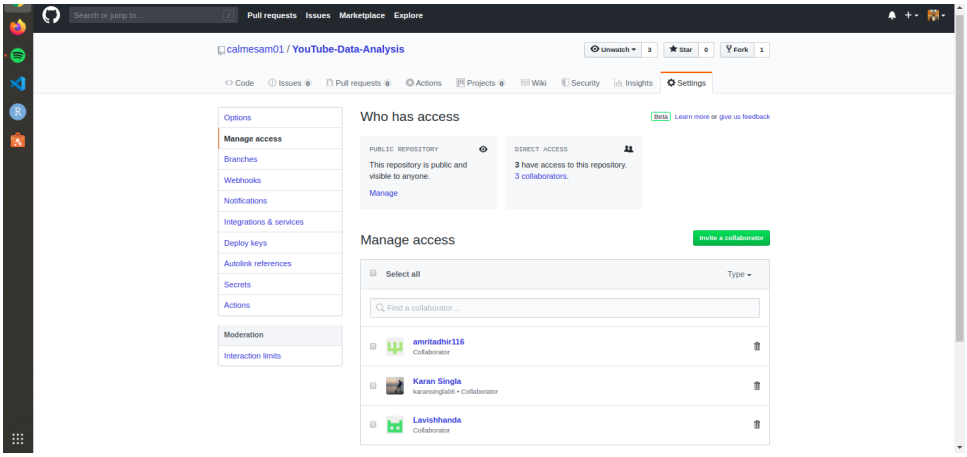
Overview

Motivation: The world's reaction to recent events like Kobe Bryant's death and coronavirus outbreak led us to analyze the YouTube trending videos. Millions of videos are posted on YouTube every single day by people in different geographical parts of the world. We want to analyze the public's reaction to such certain events and visualize the results to get a better understanding of how things are related. So far data has become the hottest technology buzzword in the last two years. The velocity of data is a factor affecting it. And the variety of data is an important set back that brings data analytics into play. The volume and velocity of data will keep on increasing as the data increases, so this data needs to be analyzed using specific tools and technologies. Moreover, data analytics is essential as it helps to present, evaluate, describe and interpret the data. Analyzing data plays a major role in terms of decision making in all the different fields of work. It involves cleaning, transforming and modelling of data. Moreover, to make any decision regarding the current scenario in any field we need to have a look at past data or future data for taking any decisions regarding it. So, everything just revolves around data analytics. Along with it keeps a check whether the data produced is enough or is being produced at an alarming rate. Now if we have a look at a different field of work from social media to health data it's just the data that influences the humans to take accurate decisions.

Short Description: Data Analytics involves investigating the data and getting the conclusions ready using specific tools and software. Data Analytics is a big field and there are various types of data analytics techniques like text, statistical, diagnostic, predictive and prospective data analysis. In the data analytics process, there should be a clear overview of the need of analyzing data and what type of analysis we have to do and also what type of data is to be collected and cleaned. After this, the data is ready for interpretation. Once the data is collected it needs to clean and set for analysis. In this project, we will be collecting YouTube's dataset, which is categorized further like likes, dislikes, number of views, number of comments, trending videos, most viewed and many more. We will be working with data analysis using tools like a spark and pyspark as these tools help in processing a large amount of data. Moreover, we will also be focusing on the data cleaning process in this project which is an important part of it that is removing all the duplicity of data and adding up the missing data. Now to collect the dataset, we will be using the YouTube API's which will provide us with the proper data set of feeds and stats related to the reviews and users.

Project Development Process

- A GitHub repository with title – YouTube-Data-Analysis, has been created.
- The team members – Sanyam, Lavish, Amrita and Karan, have been added as collaborators.



- The dataset available includes several months of data on daily trending videos for regions including US, GB, DE, CA, FR, with up to 200 videos per day.
- Several features including video_id, likes, views, comments etc are present in the dataset as shown in the image below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38916 entries, 0 to 38915
Data columns (total 16 columns):
video_id                38916 non-null object
trending_date           38916 non-null object
title                   38916 non-null object
channel_title           38916 non-null object
category_id             38916 non-null int64
publish_time            38916 non-null object
tags                    38916 non-null object
views                   38916 non-null int64
likes                   38916 non-null int64
dislikes                38916 non-null int64
comment_count           38916 non-null int64
thumbnail_link          38916 non-null object
comments_disabled       38916 non-null bool
ratings_disabled        38916 non-null bool
video_error_or_removed  38916 non-null bool
description             38304 non-null object
dtypes: bool(3), int64(5), object(8)
```

- Data transmission is in progress.
- Connection to the YouTube API, for fetching real time data, is accomplished and tested.
- Several meetings held on weekends in the Leddy Library – West and Tim Hortons.

Future plans

- The UI will be developed using Python and Jupyter notebook.
- The project will be completed by the end of March 2020.