# YouTube Trending Data Analysis

Sanyam Sareen
*University of Windsor*
*Ontario, Canada*
sareens@uwindsor.ca

Karan Singla
*University of Windsor*
*Ontario, Canada*
singlak@uwindsor.ca

*Lavish Handa*
*University of Windsor*
*Ontario, Canada*
handal@uwindsor.ca

Amrita Dhir
*University of Windsor*
*Ontario, Canada*
*dhir112@uwindsor.ca*

*Abstract* - Technology has been booming with every iteration and data generated is growing exponentially for the last few years. With new data available on the internet every now and then, expectations also rise and with expectations come insights and findings which can be used to add business value or understand customer needs. Taking the case of YouTube as an example in this project, YouTube and Google in general earn majority of their money through advertisements. The question of What Where and Whom to target might have been a problem for YouTube as well which are now being solved using data analysis techniques. As the Title of our project contains the word "Trending", it is obvious that our findings are based on the number of views videos are getting on YouTube and views mean user engagement. In this project we've drilled down the trending videos to answer most watched videos ranked under different parameters.

*Keywords—YouTube, coronavirus, trending, data, covid-19*

## I. Introduction

### A. Overview

Since the time of Internet new technologies have been evolving and impacting different businesses around the globe. Solving business problems using technology is the way to leap forward and so is this project supposed to do. The advertisement industry used to target end user is one of them and when we consider advertisement for a user base who spends most of their time on the internet, we can induce relevant information from data using data analysis techniques available to us. 2.5 Quintilian bytes of data is generated each day[1] and deriving useful insights from this enormous data is the new norm. These days corona virus is impacting the whole world and a lot of web links have come up with data which a layman user wants to see. Using data analysis techniques one can dive much deeper and get insights about this virus which are not even available on the internet. A web link can only answer questions while knowledge of data analysis can answer curiosity.

### B. Motivation

The world's reaction to recent events like Kobe Bryant's death and corona virus outbreak led us to analyze the YouTube trending videos. That's our key motivation. Millions of videos are posted on YouTube every single day by people in different geographical parts of the world. We have analyzed the public's reaction to such events and visualized the results to get a better understanding of how things are related.

### C. Significance

This project's significance will be to give meaningful insights to people who want to target certain user base and understand their habits and interests on YouTube using the Trending videos.
This project also allows to see the news channels where users most log into see the updated news of the COVID-19 and get their information for the day. These inter-linked insights are used to make decisions to add business value.

### D. Contribution

The contribution of the team members towards to various aspects of the project are given below.

- Data retrieval- Karan, Sanyam
- Virtual Environment Setup - Amrita
- YouTube Data API connection - Karan, Lavish
- Inference using Graphs- Amrita
- Data Cleaning – Sanyam, Lavish
- Handling Requests from the API - Amrita, Karan
- Version Control using Git - Sanyam, Karan
- Presentation work- Lavish, Amrita
- Bi-weekly and Final Report- Karan, Lavish
- Testing - Amrita

## II. Literature Review

The analysis of structured data has been a colossal achievement before. However, the analysis of unstructured data in the form of video format has always been a challenging task. YouTube, a Google organization, has over a billion clients and produces billions of perspectives. As YouTube data is getting created in really a huge amount and also with a great amount of speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable. The main objective of this project is to do data analysis which is a process of retrieval, cleaning, processing and modelling data to discover useful information for making business decision making. All the 4 processes are followed in doing data analysis of both the generic data and Corona Virus data. It involves data retrieval which gives a clear view about the findings and the finding based on our dataset. Ater data being collected is ready for the analysis. Now Data Cleaning process is the next onewhich involves removal of irrelevant data. The data collected may contain duplicate data or may contain errors so data cleaning removes the unwanted data. Followed by this our data is collected and cleaned and ready for analysis and analysis is done using various tools but here we have collected data from Kaggle for generic data and Corona Virus data is

collected regularly for few days analysis using YouTube API's and after the data is collected it is ready for modelling or data visualization which can be done in various using charts, graphs and text for depicting the data analysis in the project we used graphs for shows data analysis variations. Data analysis make it an ease to take decisions based on the analysis ofthe dataset.

## III. Project Details and the Methodology

### A. Definitions

Major definitions essential for the appropriate understanding of this project are mentioned below: -

- Web API: It is framework of building HTTP services that can be consumed by broad range of clientsincluding browsers, mobile phones and Tablets. It is great framework to expose data and service to different different devices. Moreover, it is an open source platform for building rest full services over.NET framework. Along with it uses the full features of HTTP like URL's request, response,headersand various content formats. Moreover, it acts as an interface between the web server or browser.Moreover, it is an web development concept having a limitation only to web application's client sideand so it does not usually include web server or any web browser application. In simple form API it acts as an interface having set of functions allowing the programmers to access specific features or data of an application, operating system or other applications. Moreover, API's can be build using various technologies like java, python, .NET.

- Matplotlib: Python has thousands of libraries and Matplolib is one of the tools used of visualization in python. It tries to make easy and hard things possible. We can generate histograms, plots, bar charts, error charts, power spectra, scatter plots etc. Moreover, this library is very flexible to use and has alotof handy and in built functions to help us out tremendously. It requires importing of data and and using of certain functions to plot the graph. [2]

- Spark: It is basically a genral purpose distributed data processing engine and is suitable for using in a broad spectrum. On top of spark there are various libraries for SQL, machine learning, graph computation and stream processing which can be put to use in an application. Moreover, it supports a lot of programming languages like java, python, Scala, R. It is capable for handling several petabytes of data at a time distributed across a cluster of thousands of cooperating virtual or physical server . [3]

- Kaggle: It is a platform for predictive data modelling and analytics competition in which companies and researchers post data and data miners fetch the data and compete to get the best model for predicting and describing the data. It also allows users to post and publish data sets and explore and build models. You can analyse different data and we can work on various data sets using Kaggle. [5]

- Python Libraries: Python libraries are a collection of functions and methods that allows you to perform without writing your code. Like PIL is one of the core libraries of python for imagemanipulation. Pillow is an actively developed fork for tis library. There are around 20 importantpython libraries used in one or other application. Moreover, the python library is an extensive curatedcollection of well documented modules.

### B. Definitions

- Matplotlib :It is an amazing visualization tool in Python library for 2D plotting of arrays. It ismultiplatform data visualization library built in NumPy arrays and designed to work with broaderSciPy Stack. We can import pyplot for plotting graph and makes it work like MATLAB. Moreoverthe visualization with it is very quick. You can use formatting style of plot using. Moreover the plotfunction is used for plotting graphs and axis function can be used to decide the axis on which graph needs to be plotted. Matplolib also allows us to pass categorial variables to many plottingfunctions .

- PySpark: Py spark is used for fast data streaming . As from log files to sensor data, application developers have to cope up with streams of data .As a data arrives in steady form from various resources. While it is certainly feasible to store data and analyse it . As it can be it can be sensible sometimes to act upon data as it arrives. So for streaming data in PY spark we upload the data by importing time stamp and string types and for streaming of data into data frame we create a data frame. Now to stream the data certain function called as write stream is used in data frame. After streaming data is ready for visualization. [4]

- Jupyter Notebook: It is a web based interactive development environment for Jupiter notebooks, code and data and also an open source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Here all data is saved as structure text files, JSON format which makes them easily shareable. It is easy to convert files in Jupiter notebooks to any format either PDF or HTML. Moreover, we can create a customized interface with it. [6]

- YouTube API: For this project we collected corona virus data using YouTube API. It basically allows developers to access video statistics and YouTube data channel using 2 types of calls that is REST and XML RCP. YouTube API for data analysis enable you to generate custom reports containing YouTube analytics data. Dimensions and Metrics are the main character of the data. Dimension basically refer to the date and time when the data was collected . Moreover each row of data has unique combination of data values . As per our choice we can choose the row and YouTube calculates the data analysis. Metrics are defined as the measurements related to user activity which includes video view counts and like and dislikes by the users. The analytics API provide us with the sorting and filtering parameters to get the accurate results. [1]

## IV. Experimental Setup

### A. Implementation

- Connecting to the YouTube Data API using PySpark.
- First the libraries including PySpark and Pandas are imported in the Jupyter Notebook.
- The dataset available includes several months of data on daily trending videos for regions including thr US, thr Great Britain, Germany, Canada and France with up to 200 videos per day.
- Dataset for Covid-19 is in the JSON format while the dataset for YouTube trending videos is in CSV format. [7]
- Several dataframes are created and they are manipulated using SQL commands.
- The bar plots and logarithmic plots help in providing inference from huge data using Seaborn and Matplotlib libraries.

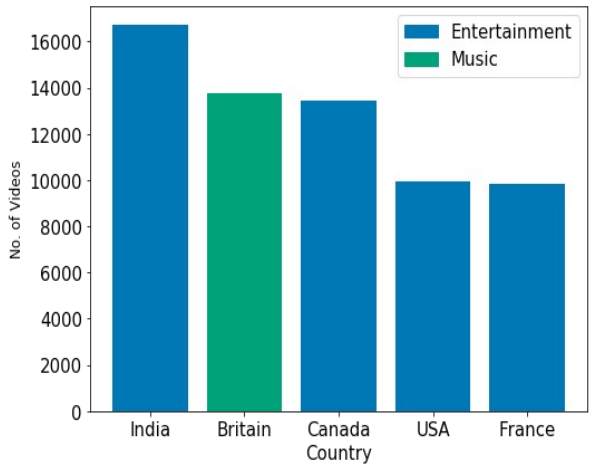


*Fig1: Most Watched Categories*

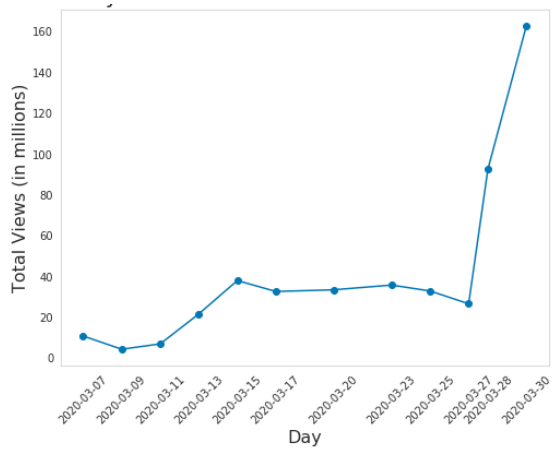


*Fig 2: Day vs total views on Coronavirus videos*

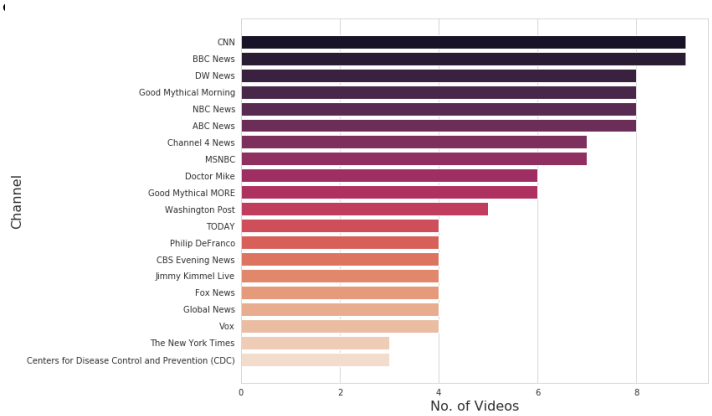


*Fig 3: Channels with most popular covid-19 videos*

### B. Testing

Performance testing is achieved and the results are following.

| Test Case ID | Test Case Description | Test Result |
|---|---|---|
| 1 | Understanding theYouTube API fetchquery | Pass |
| 2 | Responding to theYouTube APIwithcorrect response | Pass |

*Table 1: Performance testing results*

### C. Challenges

- Gathering accurate data set for corona virus analysis was a big task.
- Using SPARK technology was quite challenging in this project.
- Plotting of graphs with exact results for data visualization was quite difficult work.

### D. Limitations

- YouTube API allows only 200 entries in a single day.
- YouTube API was not feasible for collecting the count of number of deaths due to corona. [8]
- Complete data was not available as the corona virus outbreak is still going on, so we cannot generate full blown analysis.

## V. Conclusion

Data analysis on the data generated on the internet and not just on YouTube can help in getting data which can add business value to an organization looking to target their products on customers. This field has no depth and one can adjust the data to be retrieved according to their needs and is highly flexible on what to extract. Field of data analysis is expected to grow in the near future and so is the data generated by users. Trending video analysis is still an

extremely reliable source of information to target users and understand their likes and dislikes.

## VI. Future work

- Collecting More datasets.
- Working on complete analysis as Covid 19 comes to an end.
- Publish Results as a research paper.

## VII. Acknowledgment

We would like to express our gratitude towards our course mentor Dr. Pooya , who motivated us on every phase and convinced us by saying, 'You can do it'. Without his tenacious guidance, this project would not have been what it is today. In addition, a thank you to other fellow students- Saharsh Bawankar, Harsh Patel, Mandeep Singh and Neehar Arora who made our testing phase a reality.

## References

[1] YouTube Data API. Retrieved from https://developers.google.com/youtube/v3/docs/videos/list

[2] Lee, J. Introduction to Matplotlib in Python. Retrieved from https://towardsdatascience.com/introduction-to-matplotlib-in-python-5f5a9919991f

[3] Databricks. What is Spark Streaming? Retrieved from https://databricks.com/glossary/what-is-spark-streaming

[4] Shafique,A. Exploratory data Analysis using Pyspark. Retrieved from https://medium.com/@aieeshashafique/exploratory-data-analysis-using-pyspark-dataframe-in-python-bd55c02a2852

[5] Getting Started. Retrieved from https://www.kaggle.com/getting-started/44916

[6] Jupyter Notebook: An Introduction. Retrieved from https://realpython.com/jupyter-notebook-introduction/

[7] Covid-19 Coronavirus pandemic. Retrieved from https://www.worldometers.info/coronavirus/

[8] What to know about Coronavirus outbreak. Retrieved from https://www.businessinsider.com/coronavirus-in-charts-covid-19-symptoms-spread-deaths-warnings-2020-2