DSCI 551 Spring 2021 Project Final Report: Comparing Trends of Google and Twitter

**Introduction and Team Members**

Our team consists of Dhruvil Anil Trivedi and Rushi Nilesh Shah. Dhruvil Anil Trivedi is a first semester Applied Data Science student and has an undergraduate degree in Information Technology. Rushi Nilesh Shah is also first semester Applied Data Science student and has an undergraduate degree in Computer Engineering. They both have individual experiences which sums up to Web Scraping, Time Series Analysis, Data Mining, and Few of the No SQL databases. Furthermore, They both are from different fields, Rushi is from Computer Engineering which focused more on the software parts and has hands on experience on UI Development and Dhruvil is from Information Technology and has more experience in Analytical Side and has some basic knowledge about the Time Series Computation and Analysis. In addition they both are recent undergrads, so their thought process were also matching as they both had knowledge about similar technologies and frameworks. Thus they decided that by working together, they both could help each other and learn something from one another.

**Problem Description**

We are planning to compare Google Trends and Twitter Trends. Google is a humongous platform as compared to Twitter. Hence we are not comparing them directly. But on the other side, we are plan to find such topics which trend faster in google and its counterparts which trend faster on twitter. We believe there is a specific pattern to topics that trend on the google and the topics that trend on the twitter.

Google is a search engine and Twitter is social media platform. Also google’s domain is humongous if we compare it to twitter. Hence direct comparison is futile. So we plan to find the specific topics and then find their respective origin timestamp. After that, we are going to create a graph of those two platforms and find what’s the rate of spread of the trend in those two platforms with respect to that origin timestamp. In the end we believe we will be able to create a cluster of topics that trend better in google and vice versa.

For example, lets take an example of election results 2021. The first post of that in google originated at 8 November 10:58:40 And the same post originated in the twitter at 8 November 10:57:59. We plan to take a window of 30 minutes in smaller windows of 0-1,1-5,5-10,10-30, to find the rate of those trends. Assume that in first 1 minute the google had 10million searches on the election results and twitter had 1 million tweets and retweets in the first minute. We are planning to compare this with the speed of those trends in the later minutes.

**Data Collection**

Data Collection was a little difficult task as data had to be obtained from two extremely different websites, one being google and other being twitter. Google used keywords and twitter used tweets. Hence two different API’s were used for both the websites. For google, Pytrends API for python was used, it is not official API for google searches, but it was one of the most used API’s used for google searches, hence it was used. Data from 12th March 2021 to 18 March 2021 with 52 keywords was extracted using API. The task was simple for google API as compared to the twitter. Hence final google csv file was obtained directly in the end which contained the keyword and its count in normalized form for each hour.

For twitter, tweepy framework along with the official Twitter API was used. But inorder to collect all the tweets for the 52 keywords from 12th March to 18th March, premium twitter API was required. The official free twitter API only offered 100 tweets at a time. Hence snscrape package was used, it is basically a webscraping package that works for most of the social media websites. From this package we were able to obtain around 300 MB of twitter data for 6 days. This data was converted into final csv file with count of each hour. The process is explained in the next section.

**Data Preprocessing and Database Selection**

Google Data was already available in the form that was required for the project, the count of the word at each hour. But the twitter data contained tweets which contained tweets. Hence it was necessary to convert the raw twitter data into the count of tweets containing the word for each hour. For this first the data was cleaned and unnecessary attributes like username, lang, etc were removed. Then for each hour, the tweets were counted and new column was created which contained the count the count for each keyword. It is the matrix of rows representing the hour and columns representing each of the 52 keywords.

The next step was to find the Database that could help in storing the data. We looked for different big data processing databases like hive, spark, Mongodb. In the end, we decided to use spark RDD. The main columns for data is the time, keyword1, keyword2, … The main reason for using spark was the efficiency of the RDD and the ease of usage of them. The RDD can be easily queried easily, they offer the processing power of SQL and the big data capacity of Hadoop.

**Parellel Processing**

We are using Spark as the main database. The datasets are stored in using RDD formats (rows) in the Spark. RDD by itself initializes the parallel processing by creating various partitions based on the dataset inserted. The querying is also processed accordingly. The spark instance is started once and will remain on until the server is running for Django. Hence, the data processing and fetching will be executed in parallel processing fromat.

**UI/Web App Creation**

For the UI part of the project, we have decided to use the Django framework. This framework was decided because Rushi knew the framework and had worked with it in the past. Another reason for using Django was because it is MVC framework and it is easy to use. Furthermore, the Django framework was easily compatible with the Spark and it made it easier to call, handle and process the data efficiently.

**Data Science Challenge**

For the Data Science challenge, we were assigned the task created by Shree Shanmukhaa Gorla (Team 35). The task was to clean, transform and perform some basic analysis in the twitter dataset. The dataset provided was in json format and was very raw twitter data. Hence, the main focus was upon the transformation and cleaning rather than just the analytics part. The dataset was a set of json objects containing the data. The main key was data and thus the data within those keys was the actual. That data was entered into a data frame to visualize the data efficiently. The null values and subsets of dictionaries of certain entities were removed initially as they were not very consistent. As the data was very less (676 rows), there wasn’t very high analytics as compared to the preprocessing needed. Hence, we plotted a few graphs for analyzing which language and sources is used the most.

**Reflections on experiences and lessons learned**

It was a great experience throughout the project in learning and exploring various technologies. We were able to explore all tools like apache spark and frameworks like Django. We learned how to use various APIs for fetching the data and handling the datasets in a more efficient manner. We learned how to fetch, process, manage and handle data in various phases of the project and thus got a deeper understanding of the complete pipeline used in data management.

**Link to the explanation video of the presentation**

<https://drive.google.com/drive/folders/1P_OfdIbHZ51MgRw37S8ceFGg8nf5L1dz?usp=sharing>