Tweets Analysis

Ashish Patel

Gaurav Shah

Rajyash Gosalia

California State University, Los Angeles

meetashishk.patel@gmail.com
getgauravshah@gmail.com
rajyash.gosalia@gmail.com

Abstract

Big Data has been expanding and becoming one of the fundamental elements of competition, giving rise to new waves of productive growth, consumer surplus and innovation. This area has a bright future for various domains such as Google, Amazon, ebay and Yahoo. We will primarily be Focusing on collecting the tweets from the twitter website about the view of the users who are using these websites. This will help us collect the information which we will process further for carrying out analysis based on the daily tweets. Also, this data can be used by the company for analyzing and improving the area they are lacking in.

I. Introduction

In today's world online social networking and microblogging have become an integral part of the human life. Millions of users tweet and update their status online on daily basis. In September 2013, Twitter recorded that 200 million users send over 400 million tweets daily. We believe that the tweets tell us a lot about what's going on in an individual's mind. It gives us information about how the product being offered by the company is helpful and what are their personal views about that product. The tweets that we have collected are organization-specific, that is the tweets are about companies like Google, Yahoo, Amazon and eBay. The main idea behind collecting tweets about these companies is to come up with numerical analysis for each oraganization. This infomation can be useful to these big giant in various aspect such as in which region are they mostly popular and in which region/area they are lacking.

II. Methods

- Tweet Data was collected by online streaming using the Twitter API provided by the company itself.
- The collected data was cleaned and only the relevant fields were retained. Some of the field attributes that were retained are date, tweet, timezone, location and language.
- These fields were important from the analysis perspective as with these fields we could analyze the number of different languages people tweet in, tweets / second for different companies and the variation in the followers count of different companies.
- Tweets / seconds can help us analyze how many people talk about the company.
- languages can help us understand which region people tweet more about the product or the services offered by these companies.
- Followers count can help us analyze

weather there is an growth in the numbers of the people following the updates of the company.
- The above all aspects can be helpful in comparing the data and growth rate of the company in coming years.

III. Results

The collected data is Stored in the Hadoop Distributed File System(HDFS). MapReduce jobs were run on the accumulated data to come up with specific results. The results contained the final analyzed values on which visualization can be done for a better understading. One such example is displayed in the table below:-

Table 1: *Example table*

Name	
Languages	Count
en	1291
ja	2472

This is sample of the Result after collecting the tweets from all over the world. Formatted from large chunk of the tweet data which is being fetched by the Tweet API. The Result is not a single value but multiple value for the data to be fetched for analyzing. This data can be further processed for analyzing.

IV. Discussion

The analyzed data was then used to generate different graphs like the bar chart, pie chart and the scatter plot. All these graphs were plotted using the matplotlib library in python. The purpose of each of these graphs is given below:-

I. Pie Chart

Pie Chart was used to depict the percentage of tweets in each of the languages.

II. Bar Chart

Bar Chart was used to show the variation over a period of time in the followers count of the different companies.

III. Scatter Plot

Scatter plot was used to show the per second tweet of different companies.

References

[https://dev.twitter.com]
Dorsey, July 15 (2009). https://dev.twitter.com/discussions/10406