**PB**

Submitted by: (Team 11)

Moulika Chadalavada

Sampath Gattu

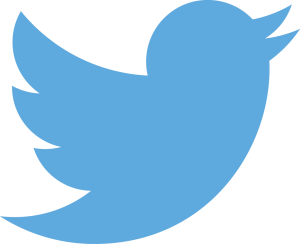
Nikitha Kona

Prathyusha

**Principles of Big Data Management**

**Phase-3**

Fall 2016



**Table of Contents**

**1.** **Introduction** 2

**1.1** **About Twitter** 2

**1.2** **Proposed Project** 2

**2.** **System Requirements** 2

**2.1** **Software Requirements** 2

**2.2** **Language Requirements** 3

**3.** **About the Theme** 3

**4.** **System Architecture** 3

**5.** **Collecting Twitter Tweets** 4

**5.1** **Generating Access Tokens** 4

**5.2** **Streaming Twitter Tweets** 4

**6.** **Analyzing Twitter Data** 4

**6.1** **Query 1: Popular Tweets on Different Diseases (Dynamically)** 4

**6.2** **Query 2: Countries that tweeted more on Diseases (Google Maps)** 6

**6.3** **Query 3: Popular Hashtags** 7

**6.4** **Query 4: Most Popular Tweeted Words (Dynamically)** 8

**6.5** **Query 5: On which day of week, more tweets are done on diseases (Dynamically)** 9

**6.6** **Query 6: Top 10 Users Tweeted on Diseases** 10

**6.7** **Query 7: Follower Id’s using Twitter API** 11

**7.** **References** 13

# **Introduction**

## **About Twitter**

Twitter is a massive social networking service that enables users to send data read short 140 character messages called ‘tweets’. More than 140 million active users publish over 400 million tweets every day. Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone and Noah Glass and launched in July 2006. Twitters speed and ease of publication have made it an important communication medium for people from all walks of life. This stream of messages from a variety of users contains information on an array of topics, including conventional new stories, events of local interest, opinions, real-time events. Twitter APIs provide access to tweets from a time range, from a user, with a keyword.

## **Proposed Project**

We choose ‘Diseases’ as our topic to do big data analysis. Based on twitter tweets, we predicted some interesting analysis on Diseases using thousands of tweets tweeted by different people. First we collected the tweets from twitter API based on some key words related to Disease. After that, we analyzed the data that we have collected. By using the analysis, we written some interesting SQL queries useful to give a proper result for the analysis.

Here we used Spark to processing the twitter data. Because it has many advantages like

* Speed: Run Programs up to 100x faster than Hadoop Map reduce in memory.
* Ease of Use: Write applications quickly in Java, Scala, Python, R.
* Generality: Combine SQL, streaming, and complex analytics.
* Runs Everywhere: Spark runs on Hadoop, Mesos, standalone, or in the cloud

# **System Requirements**

## **Software Requirements**

* Python 3.5
* JDK 1.8
* Scala 2.11.8
* Intellij IDEA
* Apache Spark
* Eclipse

## **Language Requirements**

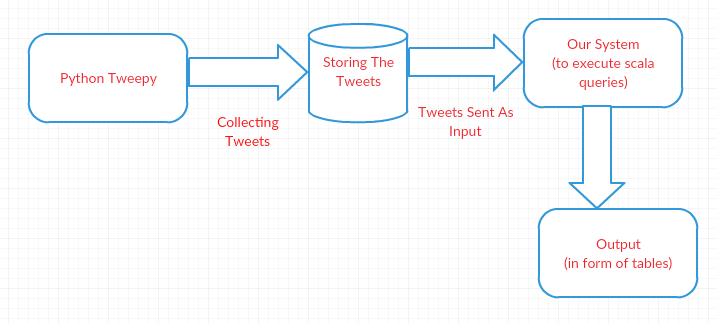
* Python
* Scala
* SQL
* HTML
* JSP
* Java
* Servlets

# **About the Theme**

We have selected the theme diseases, which provides the analysis on the twitter tweets based on the queries. Initially tweets were collected using hashtags related to disease types such as Heart Stroke, Cancer, Chicken Pox etc. Using this data, the queries are written in Scala which gives analysis on which disease more tweets were done, which user made more number of tweets on which disease, which state people in USA are more involved in tweeting about the diseases, top languages using which tweets were posted. The extracted data is displayed in tables. Based on this data visualization is done in next phase.

# **System Architecture**

First we generated credential for accessing twitter. By using these credentials, we wrote a python program to collect twitter tweets based on keywords related to food. Tweets were stored in a text file in a JSON format. We will give these JSON file to SQL queries for analysis with Spark, Intellij with Scala program with queries.



# **Collecting Twitter Tweets**

## **Generating Access Tokens**

First we generated keys for accessing twitter API. For this we need to register our application by using, [http://apps.twitter.com](http://apps.twitter.com/). We generated access token, access token secret, consumer key, consumer secret

*ACCESS\_TOKEN = "1974127951-N1QRXNCsVCywXl67BU1Wx7VlJ1fw2TlScZDY07s"*

*ACCESS\_SECRET = "zLLfZ4ZF9BxvgzjXAkjJFAVFOL4P1i0fLSaZzc6LrnqZJ"*

*CONSUMER\_KEY = "RfqrQLUtEAvwqSmNpGjZydmOn"*

*CONSUMER\_SECRET = "1XdUvxMDfcL5eE89oAo7ZO8UESv6H7HacNa9B0Y7cGFa5MiPjo"*

## **Streaming Twitter Tweets**

After that we have written a python program for streaming twitter tweets. As our theme is related to ‘Diseases’ we used few hashtags such heartstroke,cancer,malaria,braintumour,chickenpox etc. From twitter, we have streamed almost 100000 instances. Tweets were stored in JSON (JavaScript Object Notation) format in a file.

# **Analyzing Twitter Data**

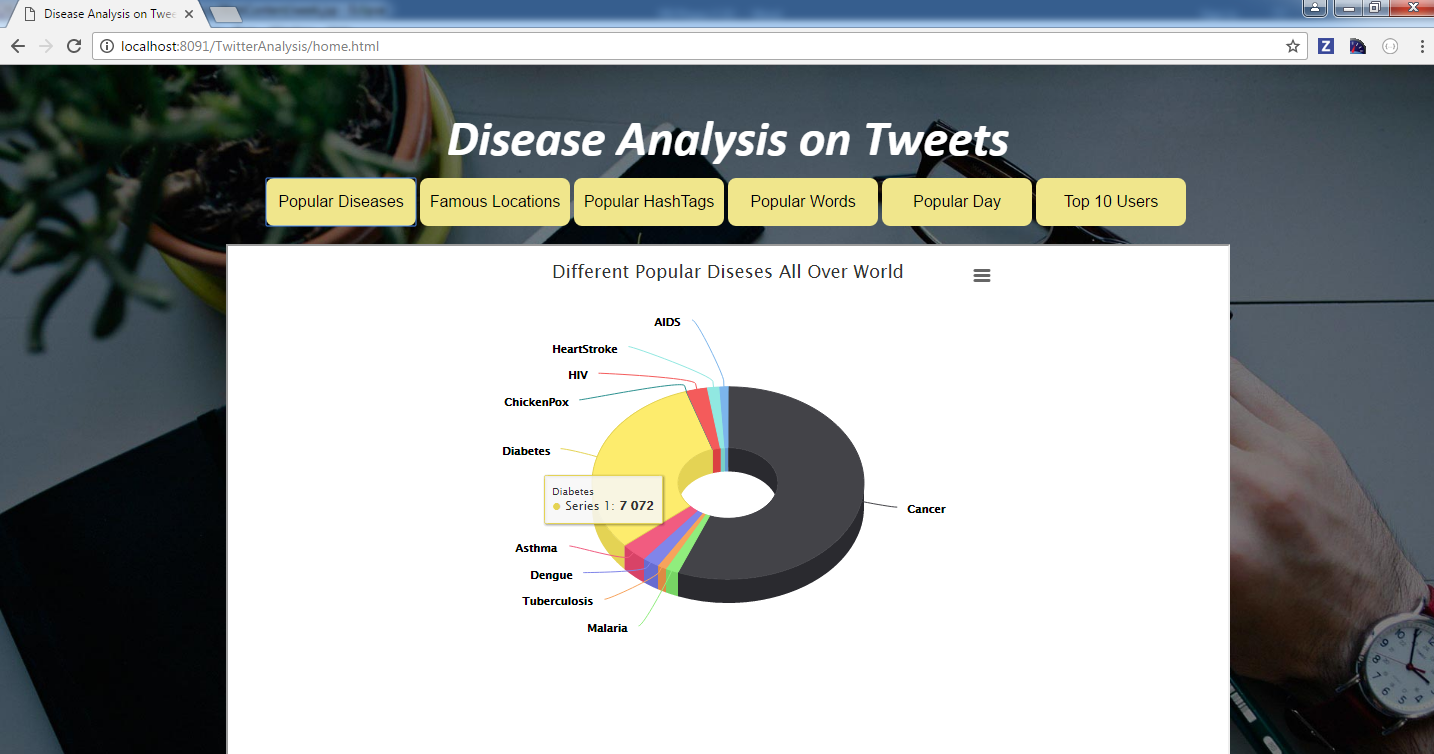
## **Query 1: Popular Tweets on Different Diseases (Dynamically)**

In this query the we are fetching maximum count of tweets done by a user on particular disease. This query is written using RDD. Initially for each disease the top tweeted user is fetched and UNION RDD is used to club all the diseases.



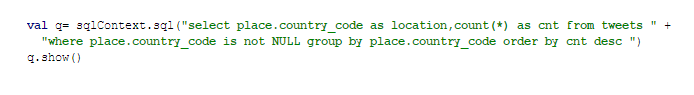


Visually, we used a PIE Chart to determine on which disease more number of tweets were posted. The visualization is dynamics though, in which we used servlet to connect front end with code and it would be able to scale the PIE Chart accordingly.



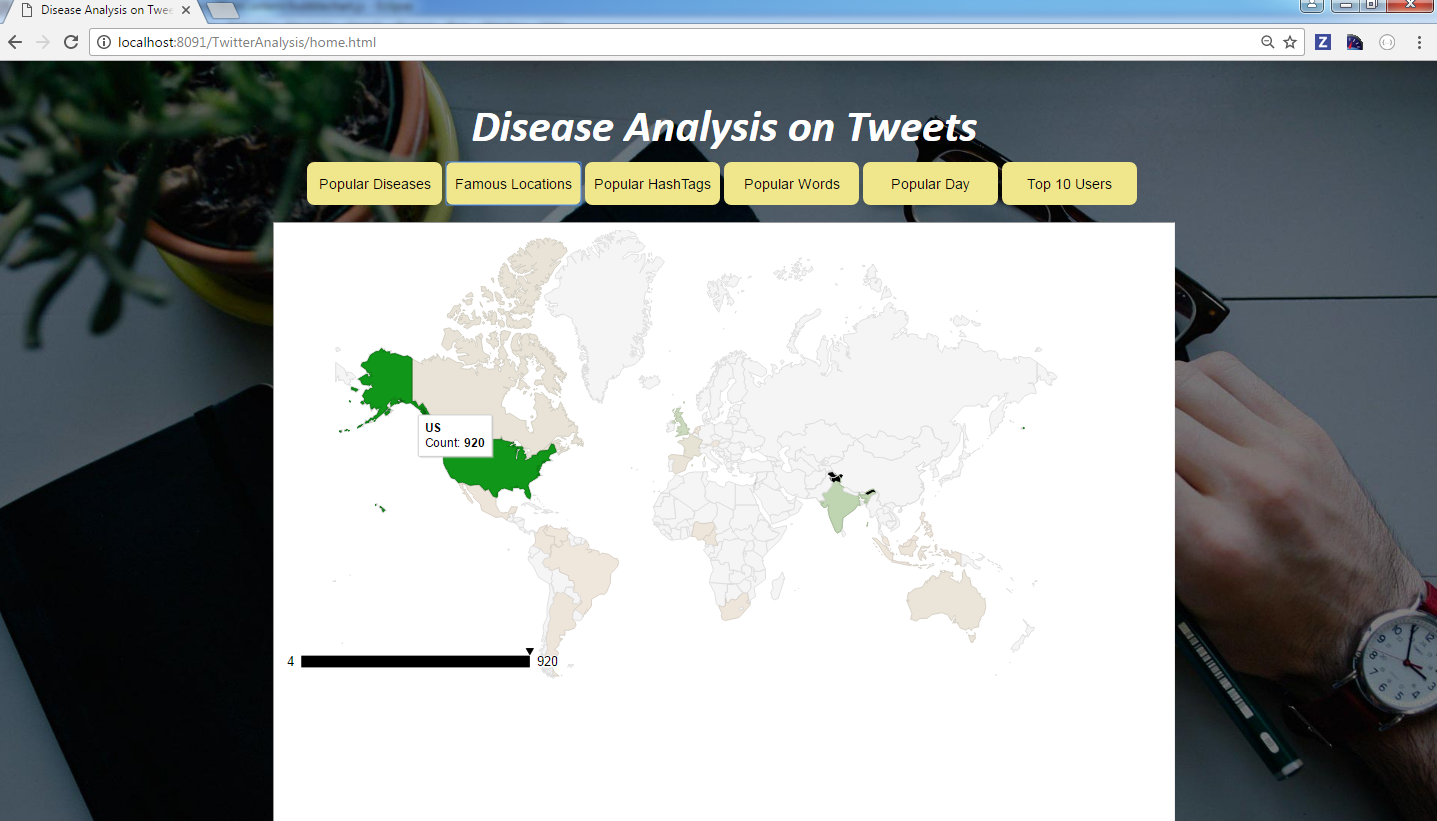
## **Query 2: Countries that tweeted more on Diseases (Google Maps)**

In this query the top countries that tweeted more on diseases is fetched. First the location in tweets are fetched from tweets file and count is displayed as shown below. The data is stored in .csv format and the file is read and Visualization is done on Google Maps.



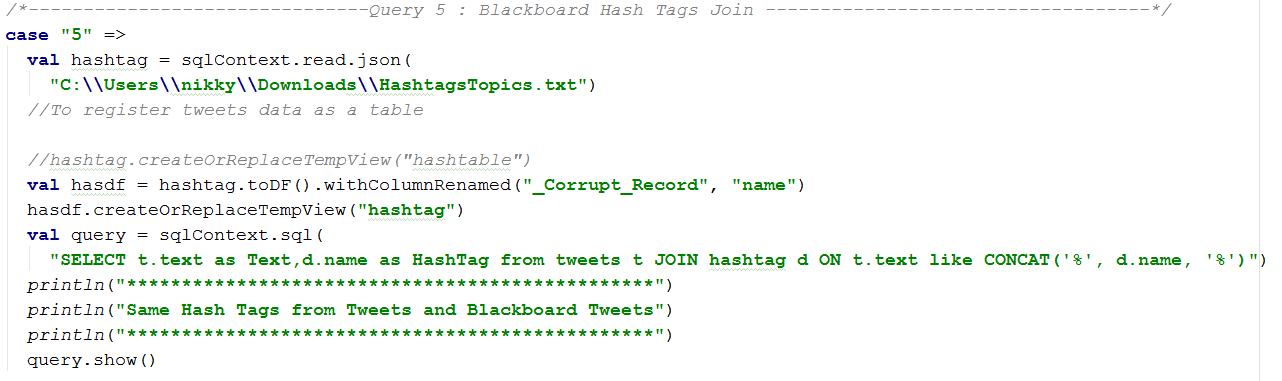
****

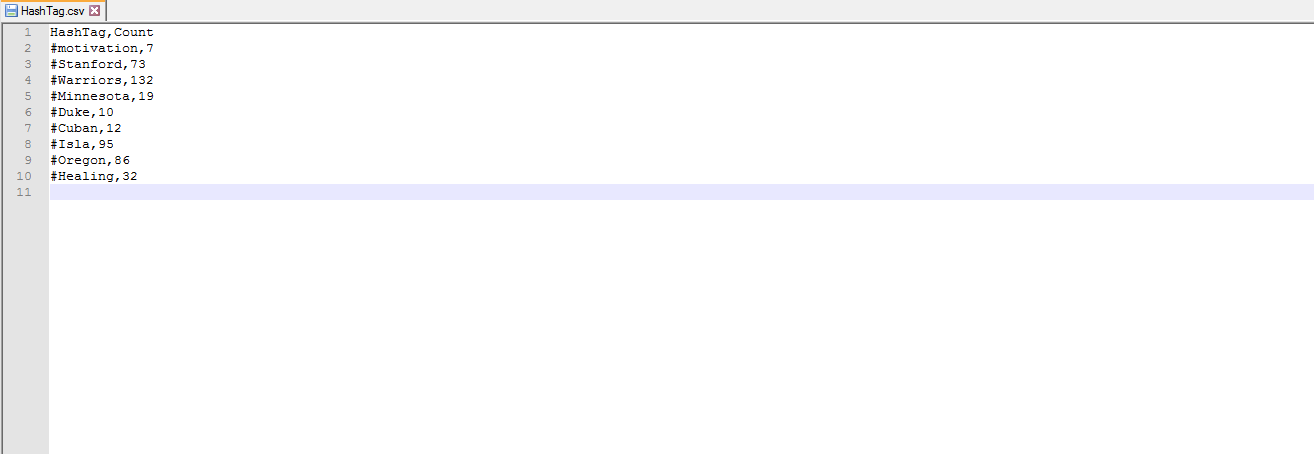
Visually, we used a google Map Chart to determine which country tweeted more number of tweets. The visualization isn’t dynamics though, in which we first run the scala code in Intellij and saved the output in a file and finally passed this file to Map chart.



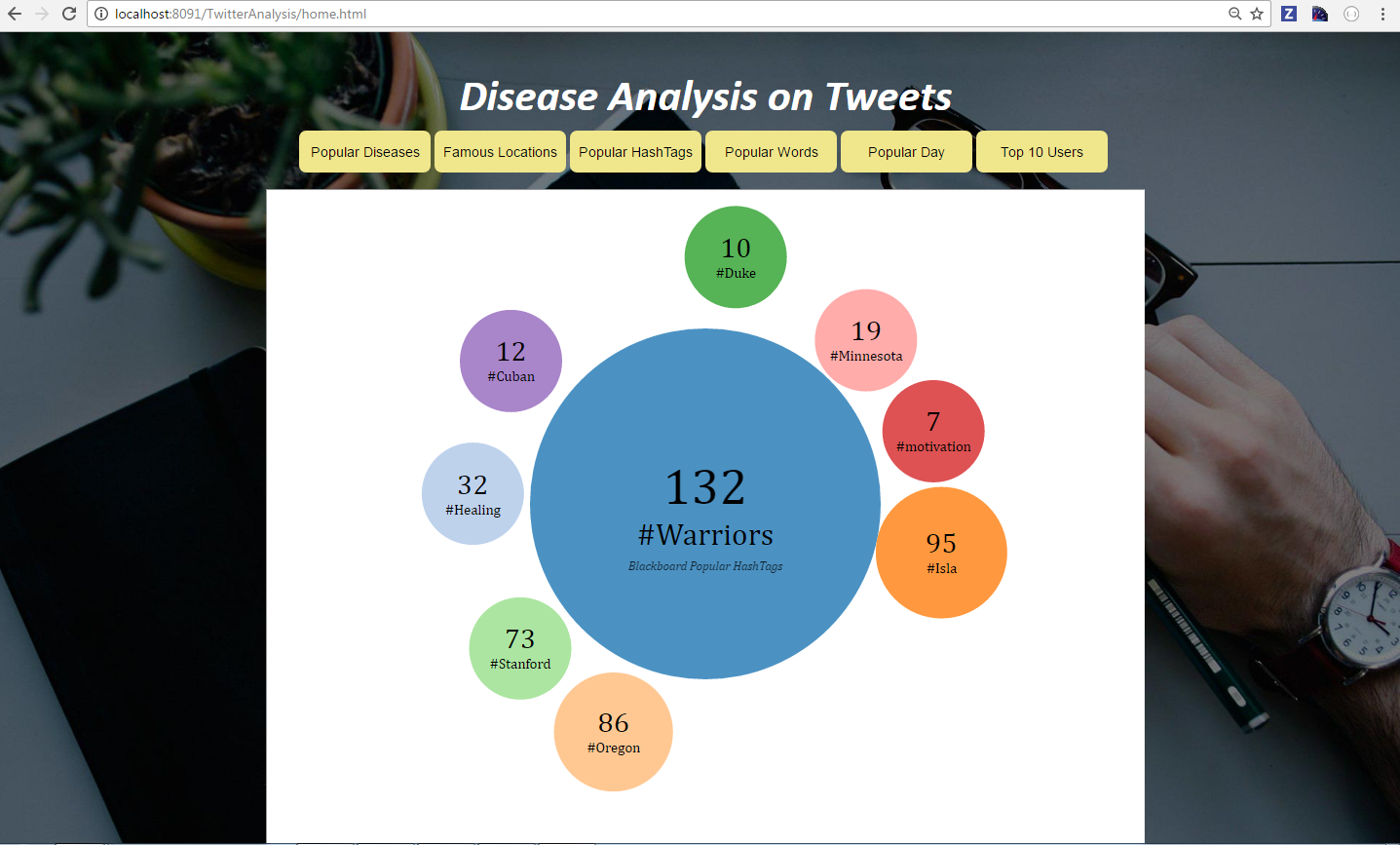
## **Query 3: Popular Hashtags**

In this query, we took popular hash tags text file from blackboard and performed JOIN operation with hash tags from diseases tweets file. The fetched data is stored in .csv format to do visualization.





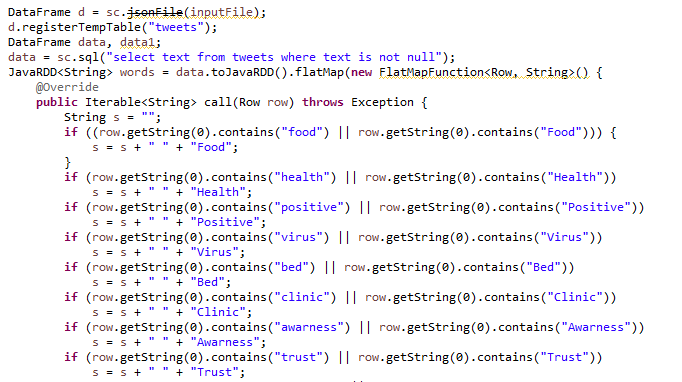
The data in csv file is used to do visualization and the representation is shown in **Bubble Chart.** According to below image **#Warriors** tag is frequently occurred HastTag in tweets file.



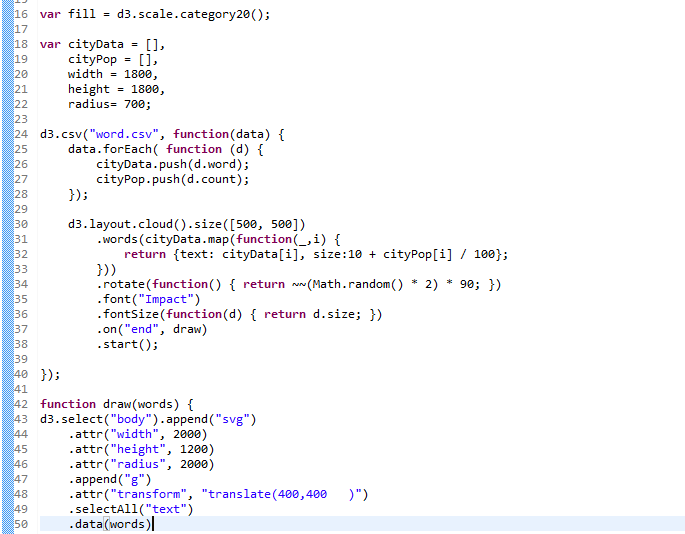
## **Query 4: Most Popular Tweeted Words (Dynamically)**

In this query, most occurring words in tweets on diseases is fetched and stored in .csv format. On the fetched data visualization is done.

**Below query is used to fetch most popular words**



**Visualization code for most popular words**



Based on the words and its counts the data is dynamically sent to wordcloud.jsp where the visualization logic is performed and below visualization is displayed on UI.

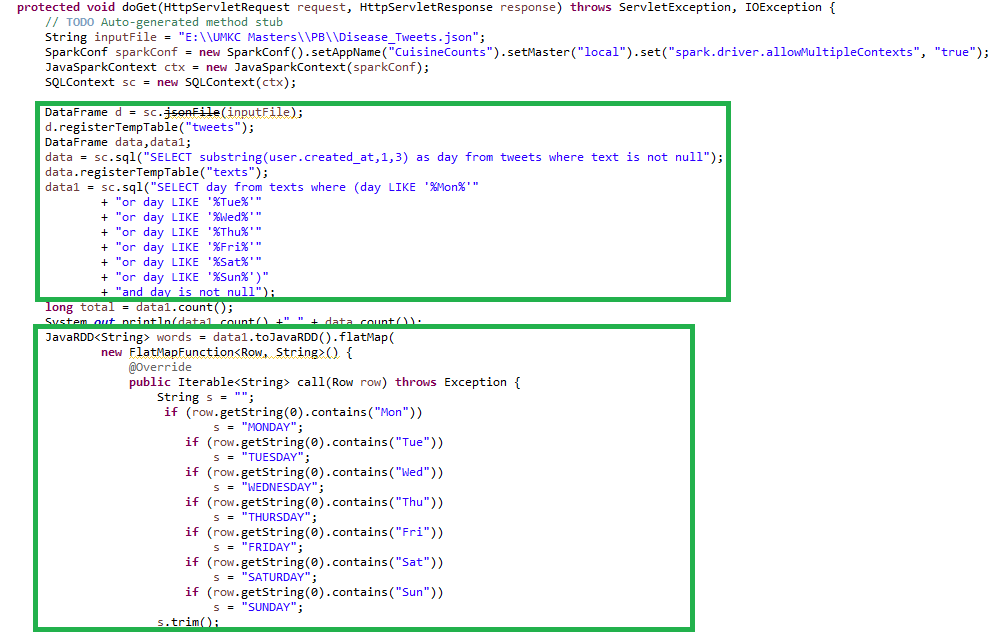


***Query Analysis: According to the query, more popular word used in tweets is Diabetes***

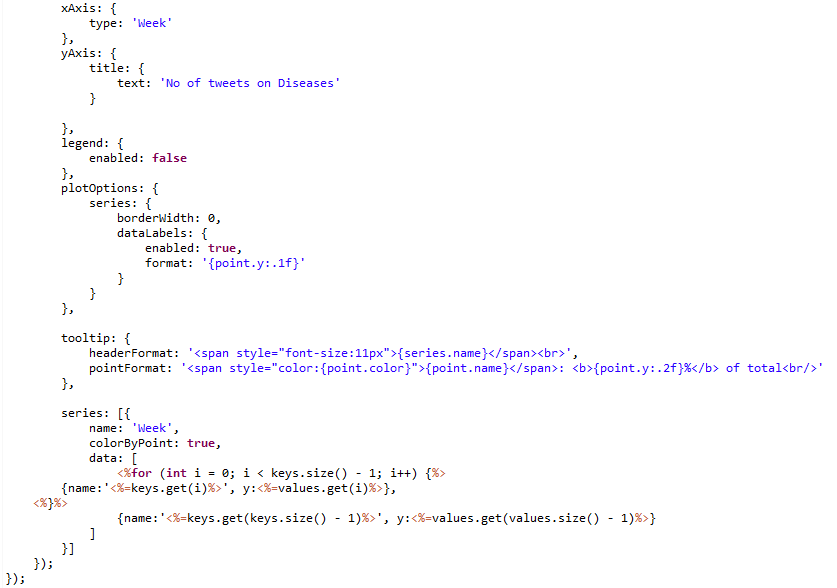
## **Query 5: On which day of week, more tweets are done on diseases (Dynamically)**

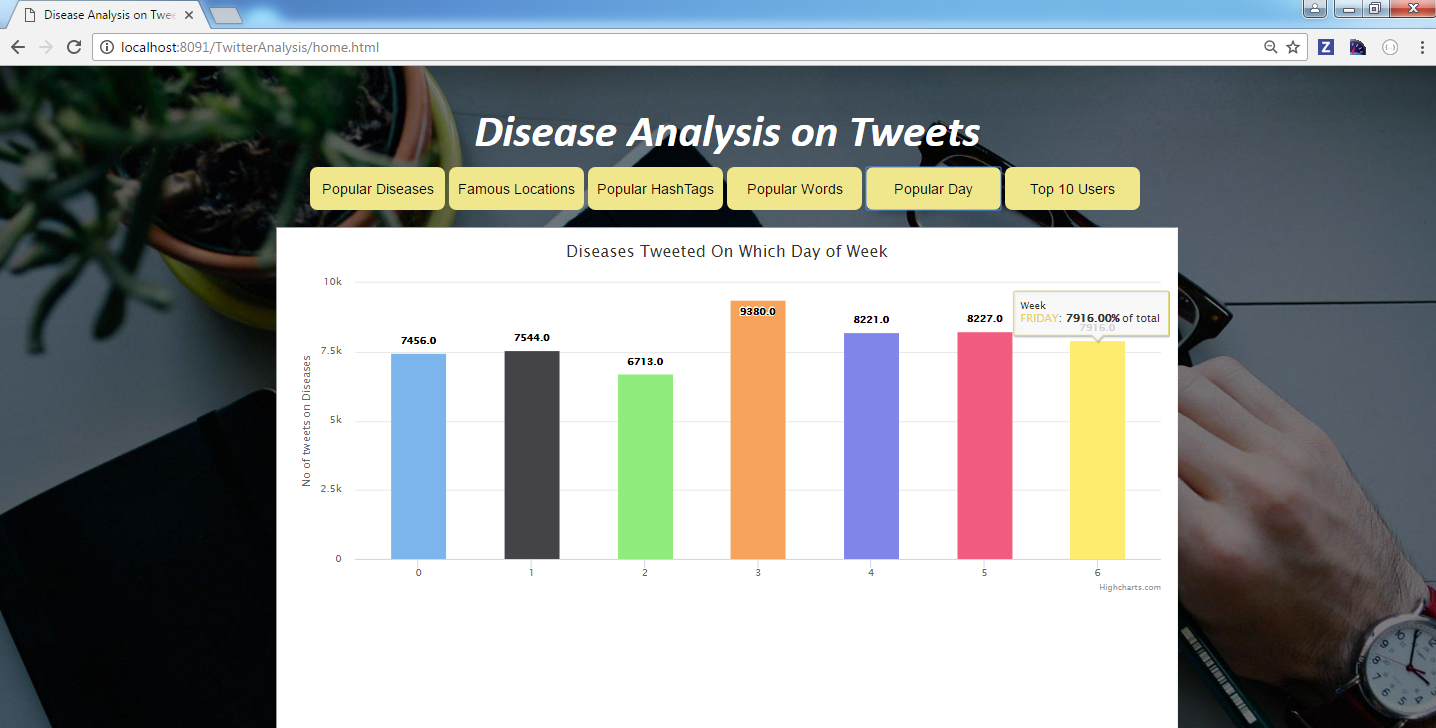
In this query, data is fetched based which day of week more tweets are done on Diseases. Initially created\_at is fetched from tweets file and count of tweets is done on each week of day.

**Java Code to fetch Week and Count from Tweets**



**JSP Code to do Visualization on Week and its count**

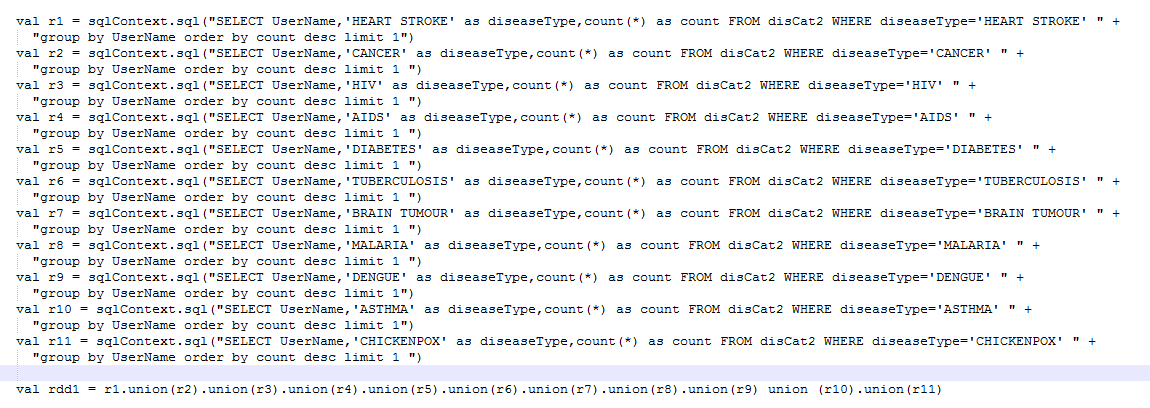


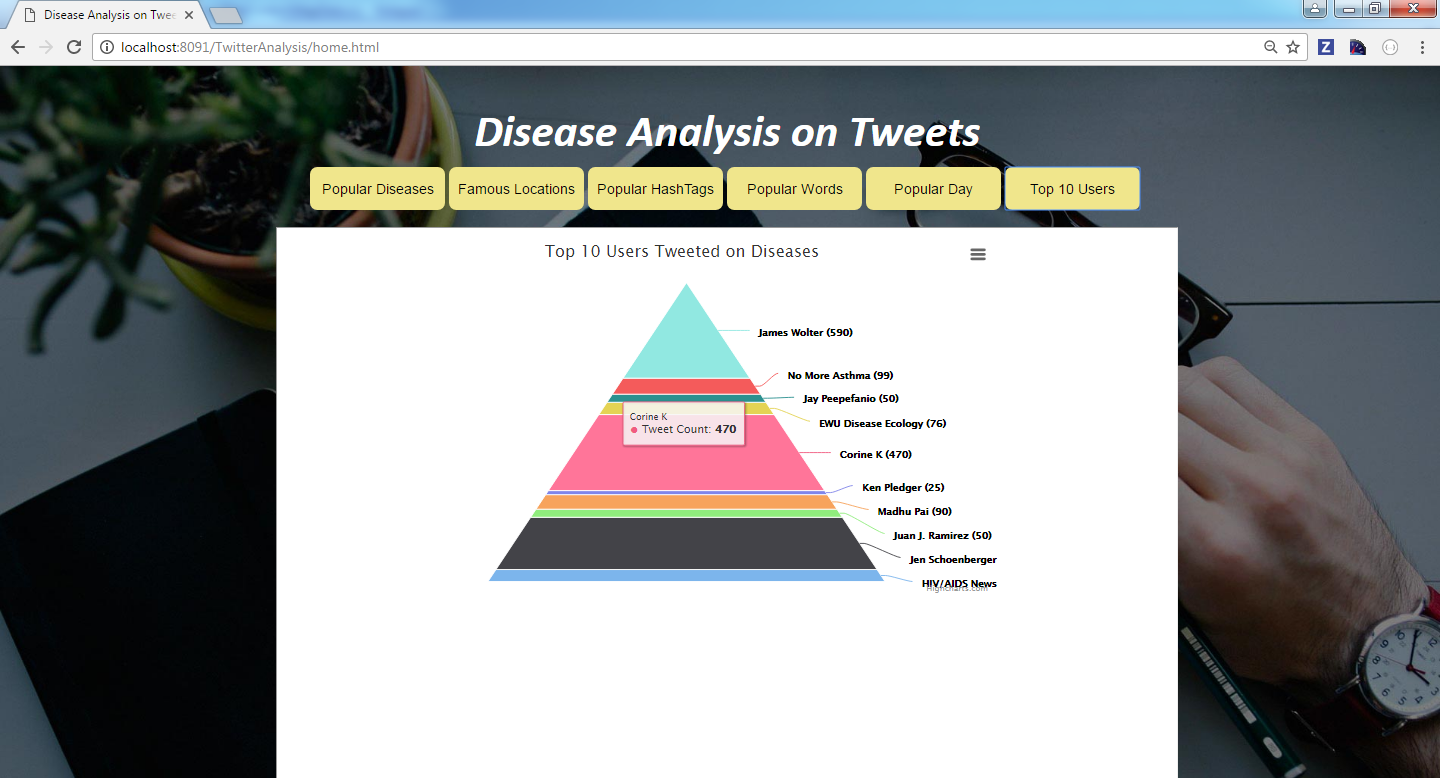


***Query Analysis: According to the query, on Thursday more tweets are done and less tweets are done on Saturday***

## **Query 6: Top 10 Users Tweeted on Diseases**

In this query the we are fetching top 10 users who tweeted more on diseases. This query is written using RDD. Initially for each disease the top tweeted user is fetched and **UNION** RDD is used to club all the diseases. The results are stored in .csv file to do visualization





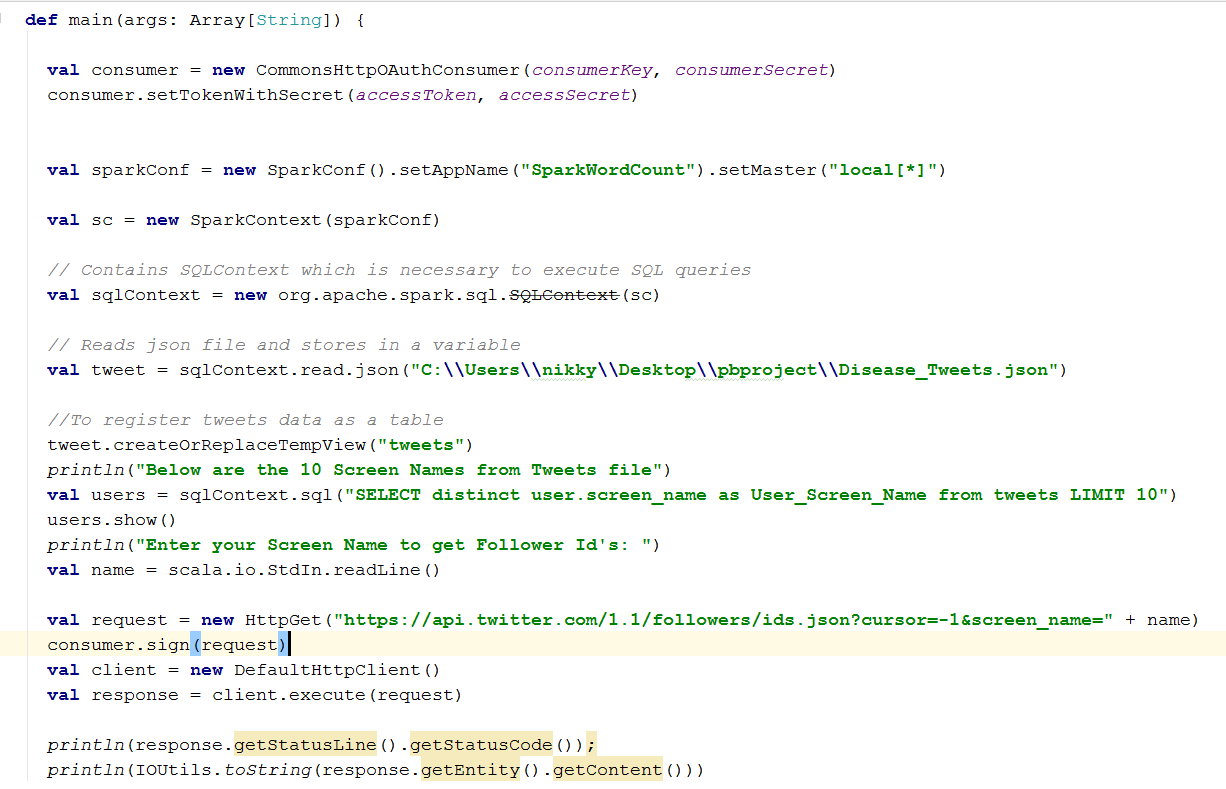
***Query Analysis: According to the query, user who tweeted more on diseases is James Wolter.***

## **Query 7: Follower Id’s using Twitter API**

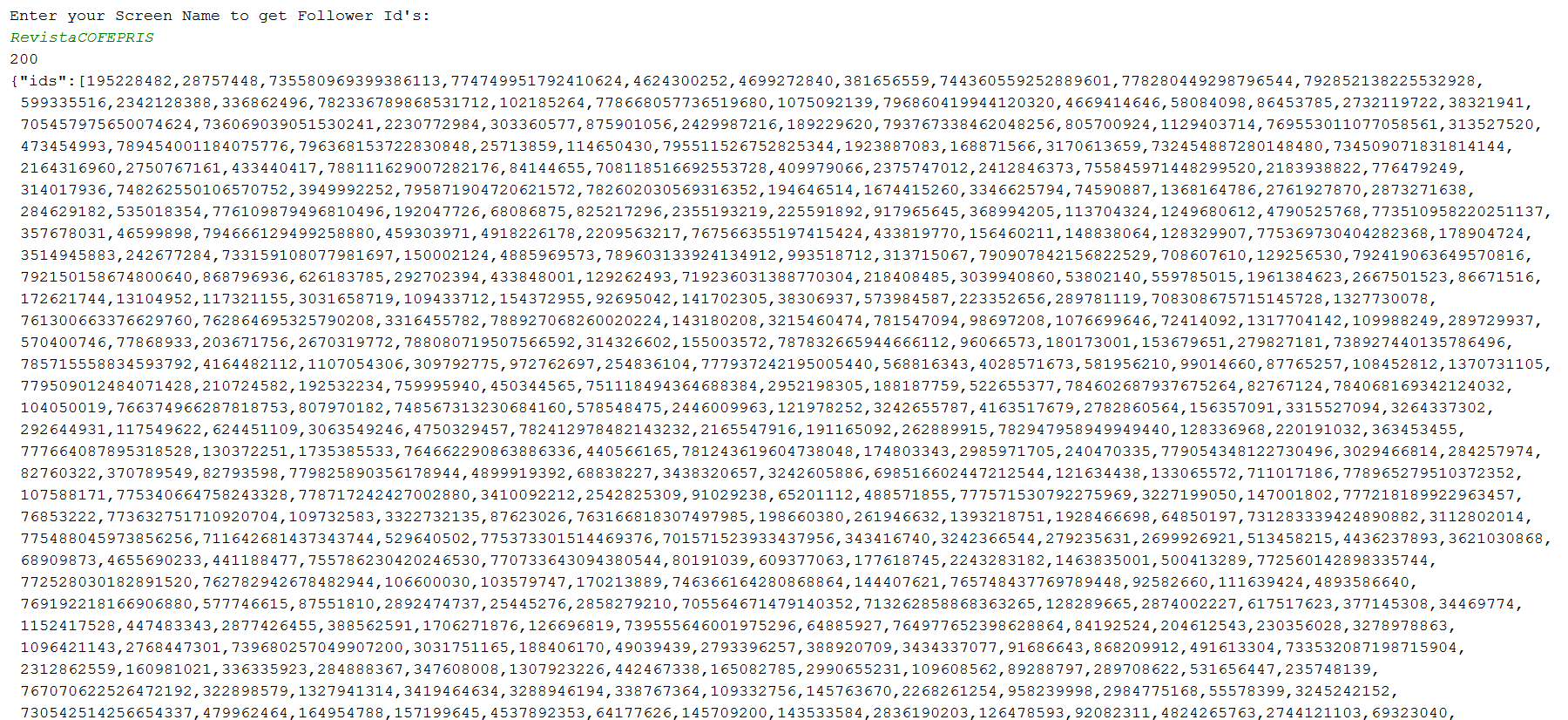
Twitter Get Followers ids API is used. A query to display ten screen names from the tweets file is written. When the query is executed a table with ten screen names is displayed in the table.

**val request = new HttpGet("https://api.twitter.com/1.1/followers/ids.json?cursor=-1&screen\_name=" + name)**

First the user is given a Choice to enter a screen name of his choice. Once the screen name has been inputted the follower’s id

******

Once screen name **RevistaCOFEPRIS** is entered the follower id’s are displayed as shown below

******

**Tweets Location:**

<https://www.dropbox.com/s/04zebrisw6jm6n0/Disease_Tweets.json?dl=0>

**Source Code:**

<https://github.com/cmoulika009/Principles-of-Big-Data-Management/tree/master/PB%20Phase-3-%20Team-11/TwitterAnalysis>

# **References**

[https://apps.twitter.com](https://apps.twitter.com/)

[http://www.tutorialspoint.com](http://www.tutorialspoint.com/)

https://www.jetbrains.com/idea/