Healthcare Analysis

Dipti Desai: dsd140330
Kaushal Shah: kbs140130
Sanket Prabhu: srp140430

Abstract:

Big data helps make the World better place. In recent years, data generation rate is growing exponentially. Every second, on average, around 6,000 tweets are tweeted on Twitter which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. The chart below shows the number of tweets per day. Big Data in healthcare is being used to predict epidemics, take preventive measures, design marketing strategies for particular pharma brands, improve quality of life and avoid preventable deaths. Social network act as a good source for providing this huge data set. Twitter will serve as a good source point to collect the data of Healthcare domain. One of the advantage of using big Data technologies is it runs on commodity hardware and as data is with no specific structure, data collected from heterogeneous sources can be stored and analyzed in a similar way. Healthcare organizations envision the future of big data. Current analytics technologies for the most part make use of discrete data and doesn't capture all types of data. Big data analysis on social networking data for Healthcare domain will make a real impact on society.

Problem Statement:

The aim of this project is to collect data from Twitter and filtering it based on a vocabulary of words related to healthcare and analyzing it to suggest brand promotion activities for pharmaceuticals, finding trending topics about Healthcare by performing clustering and performing sentiment analysis on Tweeter data.

System Overview:

To get Live Twitter data related to Healthcare. (gettweets.py)

Using tweepy:

Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication so OAuth is now the only way to use the Twitter API.
Here is a sample of how to access the Twitter API using tweepy with OAuth:

The main difference between Basic and OAuth authentication are the consumer and access keys. With Basic Authentication, it was possible to provide a username and password and access the API, but since 2010 when the Twitter started requiring OAuth, the process is a bit more complicated. An app has to be created at dev.twitter.com

OAuth is a bit more complicated initially than Basic Auth, since it requires more effort, but the benefits it offers are very lucrative:
Tweets can be customized to have a string which identifies the app which was used.
It doesn’t reveal user password, making it more secure.
It's easier to manage the permissions, for example a set of tokens and keys can be generated that only allows reading from the timelines, and so in case someone obtains those credentials, he/she won’t be able to write or send direct messages, minimizing the risk.
The application doesn't reply on a password, so even if the user changes it, the application will still work.
After logging in to the portal, and going to "Applications", a new application can be created which will provide the needed data for communicating with Twitter API.

Twitter API:
Tweepy provides access to the well documented Twitter API. With tweepy, it's possible to get any object and use any method that the official Twitter API offers. For example, a User object has its documentation athttps://dev.twitter.com/docs/platform-objects/users and following those guidelines, tweepy can get the appropriate information.
Main Model classes in the Twitter API are Tweets, Users, Entities and Places. Access to each returns a JSON-formatted response and traversing through information is very easy in Python.

Tweepy StreamingAPI:
One of the main usage cases of tweepy is monitoring for tweets and doing actions when some event happens. Key component of that is the StreamListener object, which monitors tweets in real time and catches them.
StreamListener has several methods, with on_data() and on_status() being the most useful ones. Here is a sample program which implements this behavior:

Tweepy is a great open-source library which provides access to the Twitter API for Python. Although the documentation for tweepy is a bit scarce and doesn't have many examples, the fact that it heavily relies on the Twitter API, which has excellent documentation, makes it probably the best Twitter library for Python, especially when considering the Streaming API support, which is where tweepy excels. Other libraries like python-twitter provide many functions too, but the tweepy has most active community and most commits to the code in the last year.

Clustering of live tweets related to healthcare which is obtained by running Python script.

Algorithm which we used to cluster tweets. (KMeansTweet)

Tweets Clustering using k-means

Twitter provides a service for posting short messages. In practice, many of the **health related** tweets are very similar to each other and can be clustered together. By clustering similar tweets together, we can generate a more concise and organized representation of the raw tweets, which will be very useful for Health industry for example to find trending topics to encourage research in that area.
Explanation on how to cluster tweets by utilizing Jaccard Distance metric and K-means clustering algorithm is given below.

Objectives:
· Compute the similarity between tweets using the Jaccard Distance metric.
· Cluster tweets using the K-means clustering algorithm.

Introduction to Jaccard Distance:
The Jaccard distance, which measures dissimilarity between two sample sets (A and B). It is defined as the difference of the sizes of the union and the intersection of two sets divided by the size of the union of the sets.

$$Dist(A,B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B|}$$

For example, consider the following tweets:
Tweet A: the long March
Tweet B: ides of March

|A ∩ B | = 1 and |A U B | = 5, therefore the distance is 1 – (1/5)

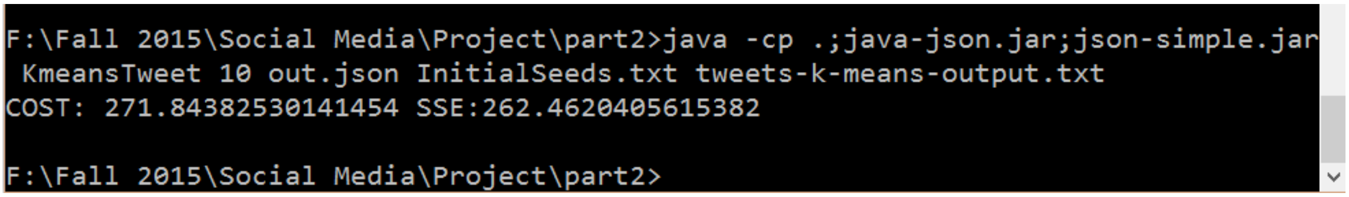
Here a tweet can be considered as an unordered set of words such as {a,b,c}.
By "unordered", we mean that {a,b,c}={b,a,c}={a,c,b}=...

A Jaccard Distance Dist (A, B) between tweet A and B has the following properties:
· It is small if tweet A and B are similar.
· It is large if they are not similar.
· It is 0 if they are the same.
· It is 1 if they are completely different (i.e., no overlapping words).

Input and obtained output for this project:

Inputs to your K-means Algorithm:

(1) The number of clusters K (for example K=25).
(2) A real world dataset related to health care sampled from Twitter by running python script.
(3) The list of initial centroids is given in the file InitialSeeds.



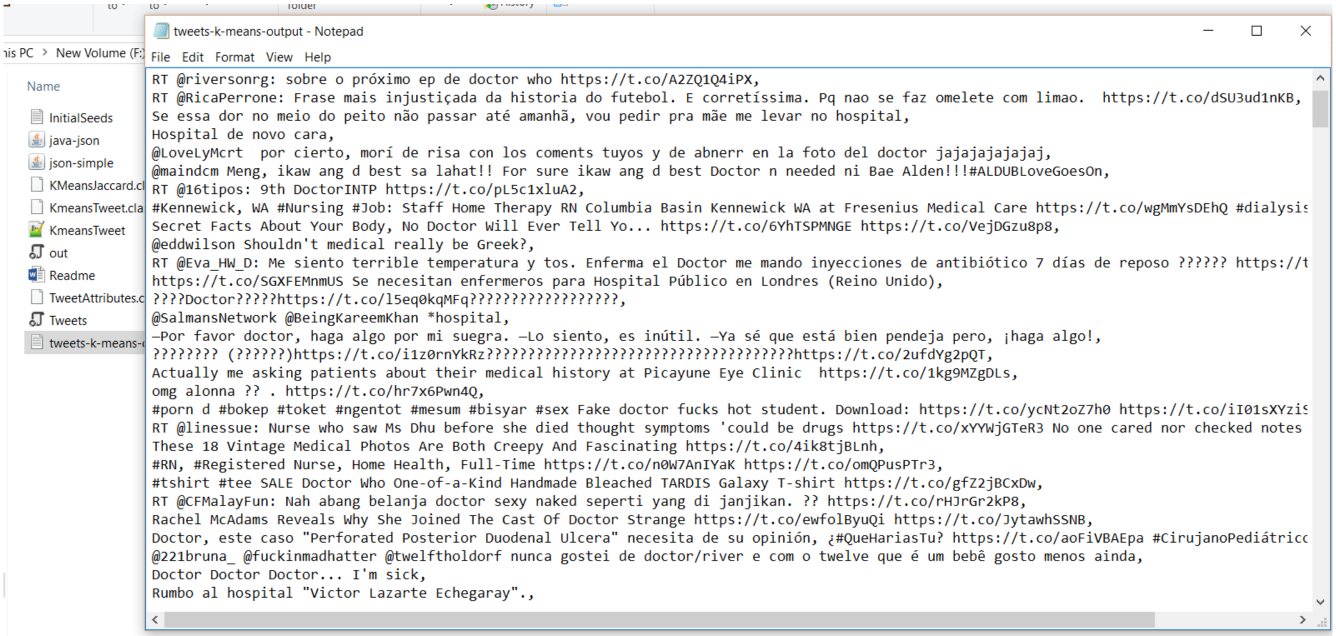
```
F:\Fall 2015\Social Media\Project\part2>java -cp .;java-json.jar;json-simple.jar
 KmeansTweet 10 out.json InitialSeeds.txt tweets-k-means-output.txt
COST: 271.84382530141454 SSE:262.4620405615382

F:\Fall 2015\Social Media\Project\part2>
```

How to run code:

- Compile kmeans java files using the following command:
  javac KmeansTweet.java -Xlint
- Run the files using following command:
  java -cp .;java-json.jar;json-simple.jar KmeansTweet 10 Tweets.json InitialSeeds.txt tweets-k-means-output.txt
  Output will get stored in tweets-k-means-output.txt file.

  Important thing to run this Program:
  I have used two external libraries one to read json and other to throw exception. I have included those two jar files.
  Names of jar files are:
- **java-json**
- **json-simple**

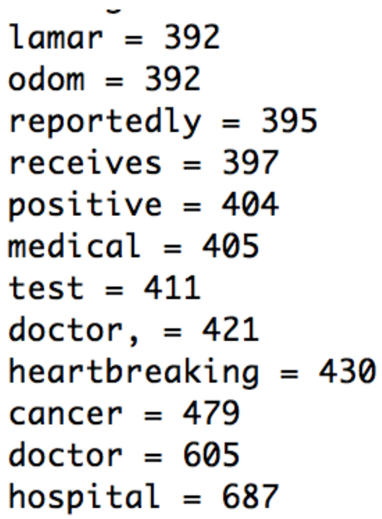
Output:



```
RT @riversonrg: sobre o próximo ep de doctor who https://t.co/A2ZQ1Q4iPX,
RT @RicaPerrone: Frase mais injustiçada da historia do futebol. E corretíssima. Pq nao se faz omelete com limao.  https://t.co/dSU3ud1nKB,
Se essa dor no meio do peito não passar até amanhã, vou pedir pra mãe me levar no hospital,
Hospital de novo cara,
@LoveLyMcrt  por cierto, morí de risa con los coments tuyos y de abnerr en la foto del doctor jajajajajajaj,
@maindcm Meng, ikaw ang d best sa lahat!! For sure ikaw ang d best Doctor n needed ni Bae Alden!!!#ALDUBLoveGoesOn,
RT @16tipos: 9th DoctorINTP https://t.co/pL5c1xluA2,
#Kennewick, WA #Nursing #Job: Staff Home Therapy RN Columbia Basin Kennewick WA at Fresenius Medical Care https://t.co/wgMmYsDEhQ #dialysis
Secret Facts About Your Body, No Doctor Will Ever Tell Yo... https://t.co/6YhTSPMNGE https://t.co/VejDGzu8p8,
@eddwilson Shouldn't medical really be Greek?,
RT @Eva_HW_D: Me siento terrible temperatura y tos. Enferma el Doctor me mando inyecciones de antibiótico 7 días de reposo ?????? https://t
https://t.co/SGXFEMnmUS Se necesitan enfermeros para Hospital Público en Londres (Reino Unido),
????Doctor?????https://t.co/l5eq0kqMFq?????????????????,
@SalmansNetwork @BeingKareemKhan *hospital,
—Por favor doctor, haga algo por mi suegra. —Lo siento, es inútil. —Ya sé que está bien pendeja pero, ¡haga algo!,
???????? (??????)https://t.co/i1z0rnYkRz????????????????????????????????????https://t.co/2ufdYg2pQT,
Actually me asking patients about their medical history at Picayune Eye Clinic  https://t.co/1kg9MZgDLs,
omg alonna ?? . https://t.co/hr7x6Pwn4Q,
#porn d #bokep #toket #ngentot #mesum #bisyar #sex Fake doctor fucks hot student. Download: https://t.co/ycNt2oZ7h0 https://t.co/iI01sXYziS
RT @linessue: Nurse who saw Ms Dhu before she died thought symptoms 'could be drugs https://t.co/xYYWjGTeR3 No one cared nor checked notes
These 18 Vintage Medical Photos Are Both Creepy And Fascinating https://t.co/4ik8tjBLnh,
#RN, #Registered Nurse, Home Health, Full-Time https://t.co/n0W7AnIYaK https://t.co/omQPusPTr3,
#tshirt #tee SALE Doctor Who One-of-a-Kind Handmade Bleached TARDIS Galaxy T-shirt https://t.co/gfZ2jBCxDw,
RT @CFMalayFun: Nah abang belanja doctor sexy naked seperti yang di janjikan. ?? https://t.co/rHJrGr2kP8,
Rachel McAdams Reveals Why She Joined The Cast Of Doctor Strange https://t.co/ewfolByuQi https://t.co/JytawhSSNB,
Doctor, este caso "Perforated Posterior Duodenal Ulcera" necesita de su opinión, ¿#QueHariasTu? https://t.co/aoFiVBAEpa #CirujanoPediátrico
@221bruna_ @fuckinmadhatter @twelftholdorf nunca gostei de doctor/river e com o twelve que é um bebê gosto menos ainda,
Doctor Doctor Doctor... I'm sick,
Rumbo al hospital "Victor Lazarte Echegaray".,
```

To find Trending Topics in Healthcare (TrendingTopics.jar)

Here we wrote **MapReduce code** which finds trending words.
(TrendingTopics.jar)



```
lamar = 392
odom = 392
reportedly = 395
receives = 397
positive = 404
medical = 405
test = 411
doctor, = 421
heartbreaking = 430
cancer = 479
doctor = 605
hospital = 687
```

Sentiment analysis on twitter data: (sentiment.py)

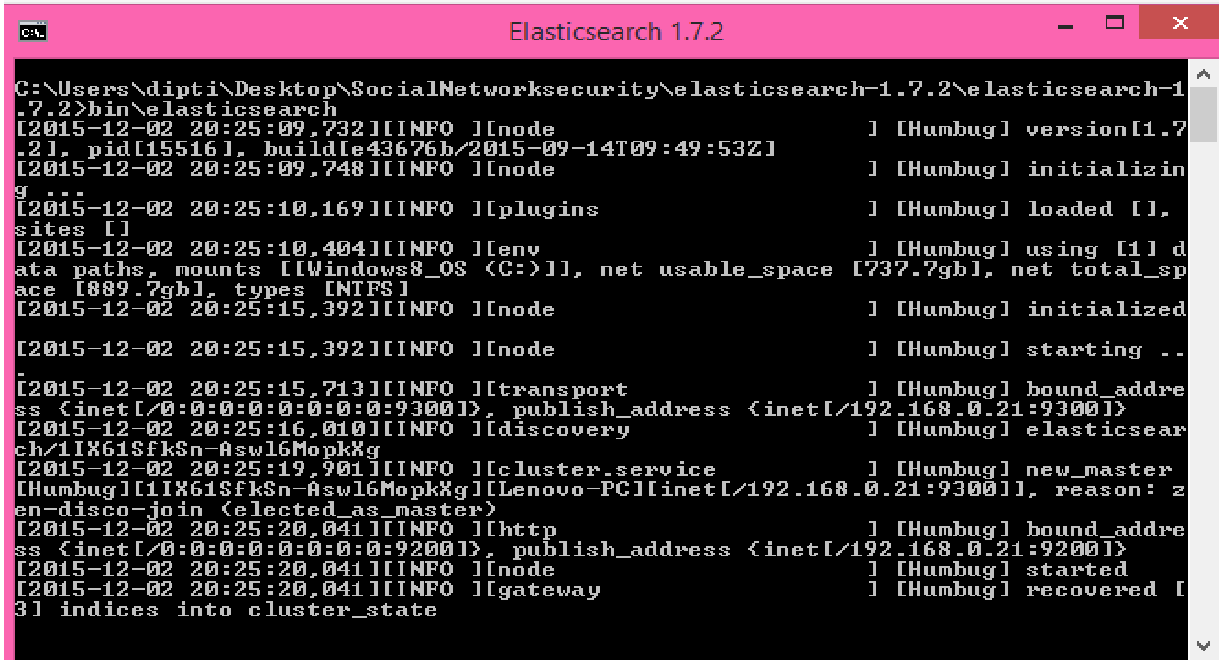
Elastic search has the ability to store large quantities of semi-structured (JSON) data and provides the ability to quickly and easily query this data. It stores data index based which makes it perfect engine to perform query and search the documents .This makes it a good option for storing Twitter data which is delivered as JSON.

Elastic search instances can be easily configured on your system.

Process to start Elastic search:

Step1:

Elastic Search can be downloaded packaged in various formats such as ZIP and TAR.GZ from elasticsearch.org. After downloading and extracting a package running it couldn't be much easier, at least if you already have a Java runtime installed.

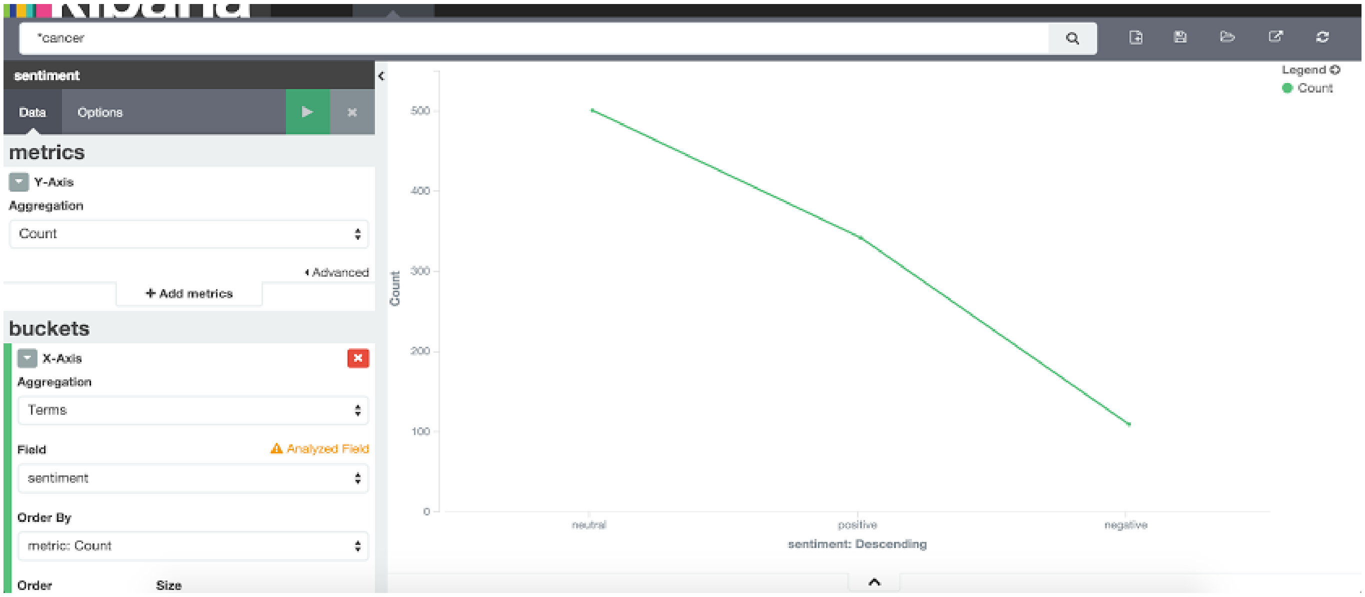


```
C:\Users\dipti\Desktop\SocialNetworksecurity\elasticsearch-1.7.2\elasticsearch-1
.7.2>bin\elasticsearch
[2015-12-02 20:25:09,732][INFO ][node                     ] [Humbug] version[1.7
.2], pid[15516], build[e43676b/2015-09-14T09:49:53Z]
[2015-12-02 20:25:09,748][INFO ][node                     ] [Humbug] initializin
g ...
[2015-12-02 20:25:10,169][INFO ][plugins                  ] [Humbug] loaded [],
sites []
[2015-12-02 20:25:10,404][INFO ][env                      ] [Humbug] using [1] d
ata paths, mounts [[Windows8_OS (C:)]], net usable_space [737.7gb], net total_sp
ace [889.7gb], types [NTFS]
[2015-12-02 20:25:15,392][INFO ][node                     ] [Humbug] initialized

[2015-12-02 20:25:15,392][INFO ][node                     ] [Humbug] starting ..
.
[2015-12-02 20:25:15,713][INFO ][transport                ] [Humbug] bound_addre
ss {inet[/0:0:0:0:0:0:0:0:9300]}, publish_address {inet[/192.168.0.21:9300]}
[2015-12-02 20:25:16,010][INFO ][discovery                ] [Humbug] elasticsear
ch/1IX61SfkSn-Aswl6MopkXg
[2015-12-02 20:25:19,901][INFO ][cluster.service          ] [Humbug] new_master
[Humbug][1IX61SfkSn-Aswl6MopkXg][Lenovo-PC][inet[/192.168.0.21:9300]], reason: z
en-disco-join (elected_as_master)
[2015-12-02 20:25:20,041][INFO ][http                     ] [Humbug] bound_addre
ss {inet[/0:0:0:0:0:0:0:0:9200]}, publish_address {inet[/192.168.0.21:9200]}
[2015-12-02 20:25:20,041][INFO ][node                     ] [Humbug] started
[2015-12-02 20:25:20,041][INFO ][gateway                  ] [Humbug] recovered [
3] indices into cluster_state
```

Once the instance is up and running, twitter data needs to be indexed and stored in the elastic search. Tweepy is one of the way to stream the real time tweets into elastic Search.
Tweepy library helps to stream real time data. Code is written in python, which stores data into elastic search and Kibana is used to visualize the data. Sentiment analysis has been performed on tweets using Text Blob. Pseudo code for the same is given below:

```
Create instance of elastic search
Connect to twitter data using Consumer key, authentication key as well as secret key
To perform sentiment analysis, refer to Text Blob
Count the polarity using sentiment value.
Create index on tweets and store in the elastic search
Before storing data, filter data for healthcare related topics.
```

Once data is stored in the elastic search, start the instance of Kibana which shows data that is indexed. Querying and searching this data becomes lot easier. Different dashboards can be created on Kibana. The data gives sentiment related to tweets. Below is the image showing sentiment analysis for cancer related tweets:



*cancer

sentiment

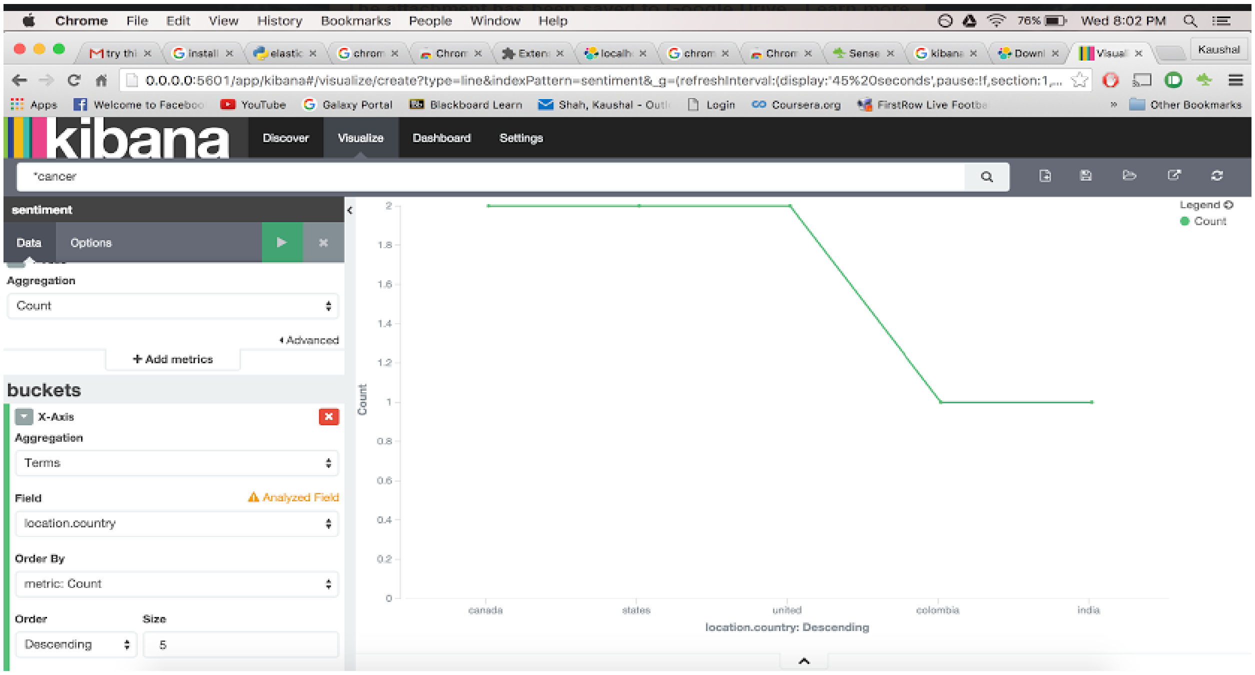
metrics

Y-Axis
Aggregation: Count

buckets

X-Axis
Aggregation: Terms
Field: sentiment
Order By: metric: Count
Order
Size

Below image shows the sentiment analysis for countries:



0.0.0.0:5601/app/kibana#/visualize/create?type=line&indexPattern=sentiment&_g=(refreshInterval:(display:'45%20seconds',pause:!f,section:1,...

*cancer

sentiment

Aggregation: Count

buckets

X-Axis

Aggregation: Terms

Field: location.country

Order By: metric: Count

Order: Descending

Size: 5

Software/Tools?

- Python IDE
- Logstash
- Hadoop
- Java (JDK)
- Elastic Search (NOSQL DATABASE)
- KIBANA (making data easy to understand!!!)

References:

[1] http://www.internetlivestats.com/twitter-statistics/
[2] https://www.healthcatalyst.com/big-data-in-healthcare-made-simple
[3] https://dev.twitter.com/rest/public
[4] https://www.elastic.co/products/hadoop
[5]http://www.rittmanmead.com/2015/08/three-easy-ways-to-stream-twitter-data-into-elasticsearch/