Twitter Spam Detection & Sentimental Analysis

Submitted by: Praneeth Krishna, Sakshi Arora, Prateek Gangwal

Under the Guidance of

Prof. Srikanth Krishnamurthy &

TA Spandan Brahmbhatt

Table of Contents

Environment setup

For our project we need below software setup

- Intellij IDEA 4.3
- Apache Spark 1.4.0
- Elastic search server 1.4.4
- Kibana 4.0.1

Intellij IDEA 4.3

- Download latest version on Intellij Idea 4.3 from https://www.jetbrains.com/idea/download/ and install scala and sbt plugins Apache Spark 1.4.0
- Add spark dependencies in built.sbt file in your project directory and download respective .jar files
    - libraryDependencies += "org.apache.spark" %% "spark-core" % "1.4.0"
    - libraryDependencies += "org.apache.spark" %% "spark-mllib" % "1.4.0"
    - libraryDependencies += "org.apache.spark" %% "spark-streaming" % "1.4.0"
    - libraryDependencies += "org.apache.spark" %% "spark-streaming-twitter" % "1.4.0"

Elastic Search Server 1.4.4

- We need to setup one t2.large Amazon Linux AMI machine with 8gb memory and 25gb disk space.
- Choose security rules wide open to make free communication between kibana and spark local instance.
- Follow below steps for installing elastic search server.
    - sudo su
    - yum update -y
    - cd /root
    - wget https://download.elasticsearch.org/elasticsearch/elasticsearch/elasticsearch-1.4.4.noarch.rpm

```
▪ yum install elasticsearch-1.4.4.noarch.rpm -y
▪ rm -f elasticsearch-1.4.4.noarch.rpm
▪ cd /usr/share/elasticsearch/
▪ ./bin/plugin -install mobz/elasticsearch-head
▪ ./bin/plugin -install lukas-vlcek/bigdesk
▪ ./bin/plugin install elasticsearch/elasticsearch-cloud-aws/2.4.1
▪ cd /etc/elasticsearch
▪ nano elasticsearch.yml
```

- We need to change the config file as follows

```
cluster.name: Twitter_Streaming

cloud.aws.access_key: *******

cloud.aws.secret_key:  *******

cloud.aws.region: us-east-1

discovery.type: ec2

discovery.ec2.tag.Name: "Twitter_Streaming_ElasticSearch"

http.cors.enabled: true
http.cors.allow-origin: "*"
```

- Start elastic search server using command: service elasticsearch start and you can access elastic search url from localhost:9200

Kibana 4.0.1

- We need to setup one t2.large Amazon Linux AMI machine with 8gb memory and 25gb disk space.
- Choose security rules wide open to make free communication between elastic search and spark local instance.

- Follow below steps for installing kibana server.

```
sudo su

yum update -y

cd /root

wget https://download.elasticsearch.org/kibana/kibana/kibana-4.0.1-linux-x64.tar.gz

tar xzf kibana-4.0.1-linux-x64.tar.gz

rm -f kibana-4.0.1-linux-x64.tar.gz

cd kibana-4.0.1-linux-x64

nano config/kibana.yml
```

Config

```
elasticsearch_url: "ELASTICSEARCH_URL_HERE"
```

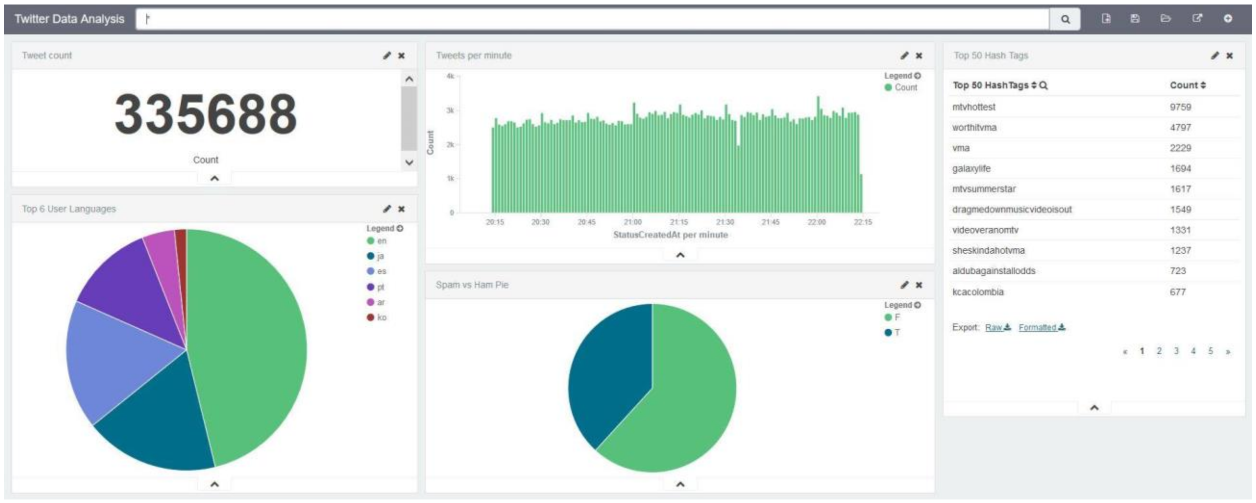
Commands

```
nohup ./bin/kibana &
```

Twitter Statistics

- 645,750,000 registered users.
- 135,000 new users signing up every day.
- 500 million tweets are generated every day.
- 9100 tweets per second.

Twitter Statistics – Kibana Representation

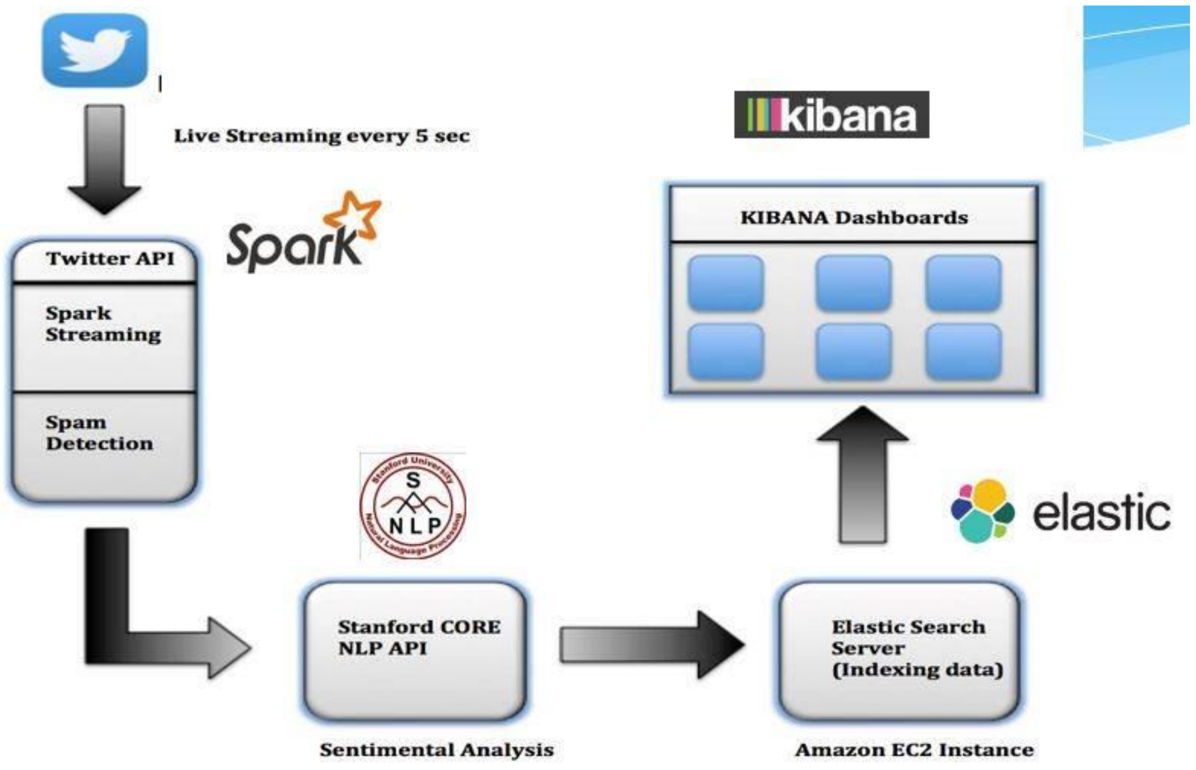


Top 50 HashTags	Count
mtvhottest	9759
worthitvma	4797
vma	2229
galaxylife	1694
mtvsummerstar	1617
dragmedownmusicvideoisout	1549
videoveranomtv	1331
sheskindahotvma	1237
aldubagainstallodds	723
kcacolombia	677

Dataset

- Spark Streaming is used to collect tweets as the dataset. The tweets are written out in JSON format, one tweet per line.
- TwitterUtils in the Spark Streaming Twitter library to get a DStream of tweets.

Workflow Architecture



Why this technology Stack?

	Storm	**Spark Streaming**
Processing Model	Record at a time	Mini batches
Latency	Sub-second	Few Second
Fault tolerance- every record processed	At least once (may be duplicates)	Exactly once

- Spark applications can interact with Elasticsearch just as they would with an HDFS resource, allowing them to index and analyze data transparently, in real-time.
- Elasticsearch acts as a database to store the tweet information.
- Kibana is used to explore massive amounts of data in Elasticsearch through easy-to-generate reports in real time.

Spam Detection

Spam detection function

To get the best results it's important that you filter to content that is relevant to your use case, and to remove what you consider as spam. As the question 'What is spam?' is so use-case specific, it's impossible to write a generic spam filter for all applications. Instead, here we'll look at ideas you can apply to your streams.

Building an ideal spam filter:

- Recently Created Users
  A common tactic used by spam bots is to create new accounts on-the-fly and tweet from these accounts until they are shut down. You can use the twitter profile age to see how old the profile is (in days), and filter out content from recently created users.

```
Days.daysBetween(new
DateTime(formatter.parse(tweet.get("UserCreated").mkString).getTime),
  DateTime.now).getDays < 1
```

- Users That Create Little Content
  Similar to inspecting a user's profile age, if a user has hardly ever tweeted it may be that the user was created by a bot to deliver a small amount of tweets before it is discarded

```
tweet.get("Text").mkString.split(" ").filter(_.startsWith("#")).length < 50
```

- Users With Few Followers
  If a Twitter profile is created just to post spam messages, the profile is likely to follow lots of users, but be followed by very few users itself. If this is the case then the users followers ratio (number of users who follow the user, divided by the number of users they follow) will be low

```
("UserFollowersRatio" -> status.getUser.getFollowersCount.toFloat /
status.getUser.getFriendsCount.toFloat)
```

- Users With Short Descriptions
  If a Twitter profile has no description, again it could be from a bot or at least from a user who doesn't care about their public profile and has little concern for the quality of content they post.

```
tweet.get("UserDescription").mkString.length < 20
```

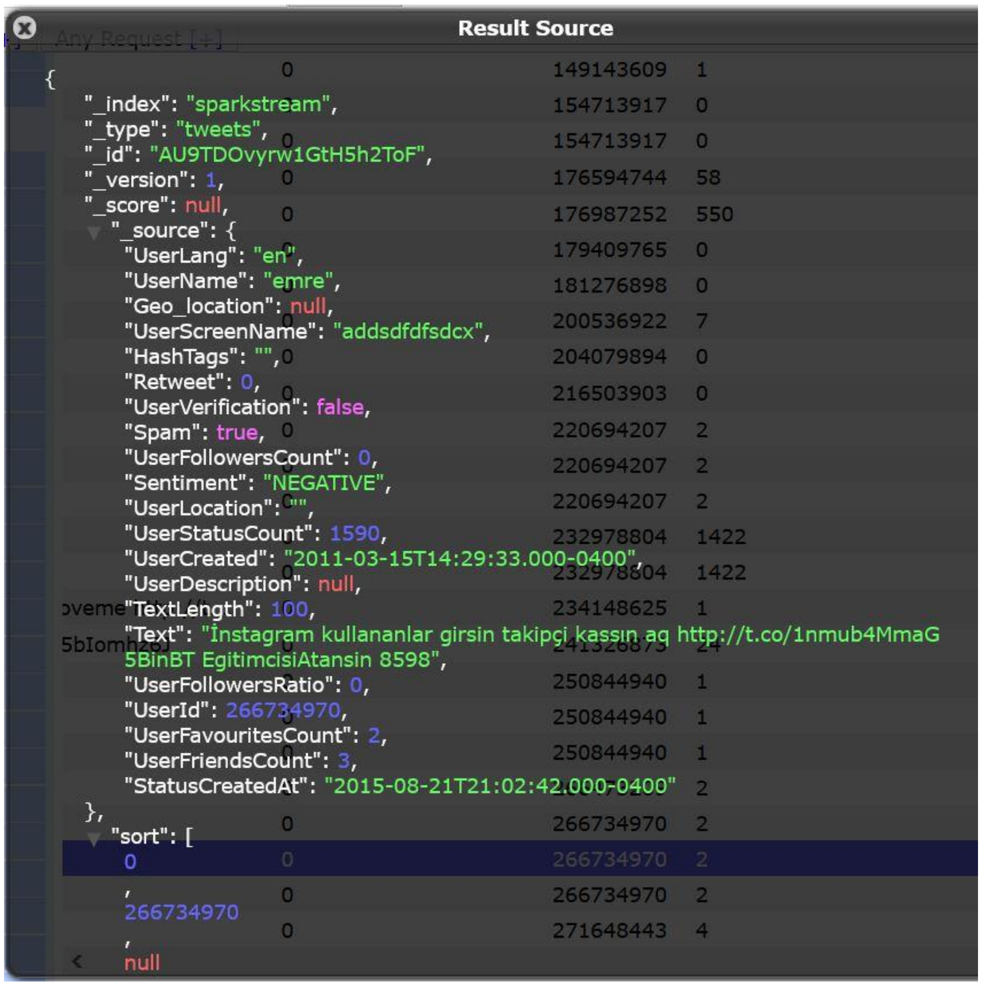
- Large Numbers Of HashTags
  Poor quality content tends to include many hashtags. Many hashtags might be used by spam creators to hope that they can reach as many users listening to those tags as possible

```
tweet.get("Text").mkString.split(" ").filter(_.startsWith("#")).length > 5
```

- Short Content Length
  Often users will write very short posts, such as '@friend ok' as a response to a question. This conent has little value in analysis.

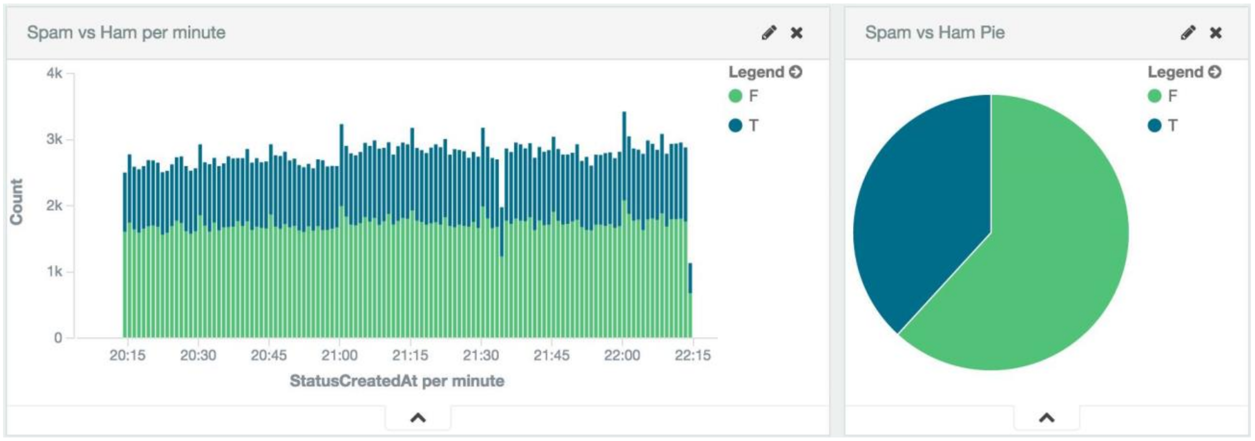
```
tweet.get("TextLength").mkString.toInt < 20
```

A spam user Example Screenshot


Result Source

```
{
  "_index": "sparkstream",
  "_type": "tweets",
  "_id": "AU9TDOvyrw1GtH5h2ToF",
  "_version": 1,
  "_score": null,
  "_source": {
    "UserLang": "en",
    "UserName": "emre",
    "Geo_location": null,
    "UserScreenName": "addsdfdfsdcx",
    "HashTags": "",
    "Retweet": 0,
    "UserVerification": false,
    "Spam": true,
    "UserFollowersCount": 0,
    "Sentiment": "NEGATIVE",
    "UserLocation": "",
    "UserStatusCount": 1590,
    "UserCreated": "2011-03-15T14:29:33.000-0400",
    "UserDescription": null,
    "TextLength": 100,
    "Text": "İnstagram kullananlar girsin takipçi kassın aq http://t.co/1nmub4MmaG
    5BinBT EgitimcisiAtansin 8598",
    "UserFollowersRatio": 0,
    "UserId": 266734970,
    "UserFavouritesCount": 2,
    "UserFriendsCount": 3,
    "StatusCreatedAt": "2015-08-21T21:02:42.000-0400"
  },
  "sort": [
    0
    ,
    266734970
    ,
    null
```

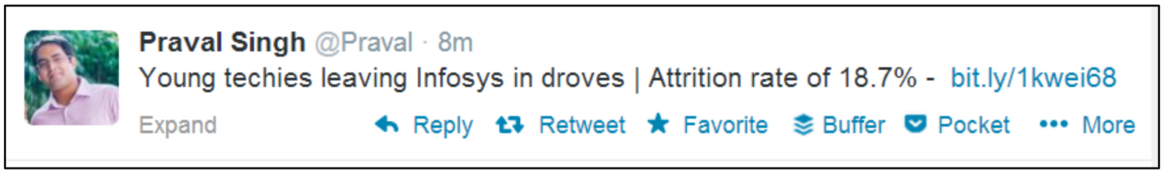
Spam Vs Ham – Kibana Representation



Sentiment Analysis

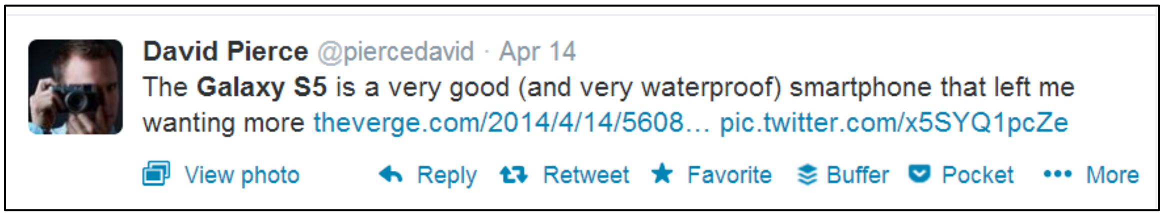
What is sentiment analysis?

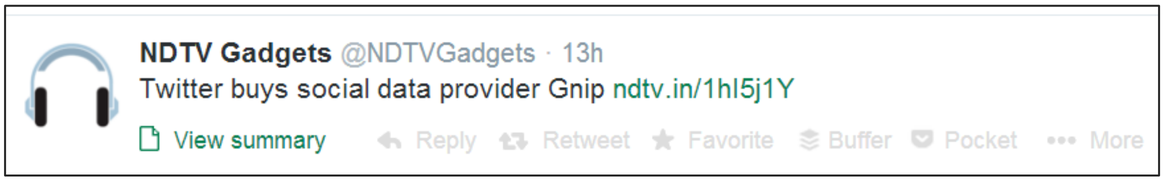
- It's a classification of polarity of a given text in the document, sentence or phrase
- The goal is to determine whether the expressed opinion in the text is positive, negative or neutral


Praval Singh @Praval · 8m
Young techies leaving Infosys in droves | Attrition rate of 18.7% - bit.ly/1kwei68

Negative

Positive


David Pierce @piercedavid · Apr 14
The **Galaxy S5** is a very good (and very waterproof) smartphone that left me wanting more theverge.com/2014/4/14/5608… pic.twitter.com/x5SYQ1pcZe


NDTV Gadgets @NDTVGadgets · 13h
Twitter buys social data provider Gnip ndtv.in/1hI5j1Y

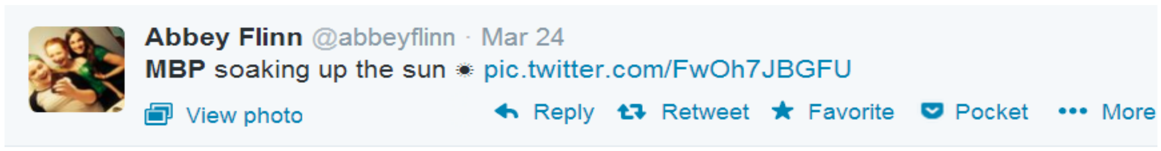
Neutral

Why is Sentiment Analysis Important?

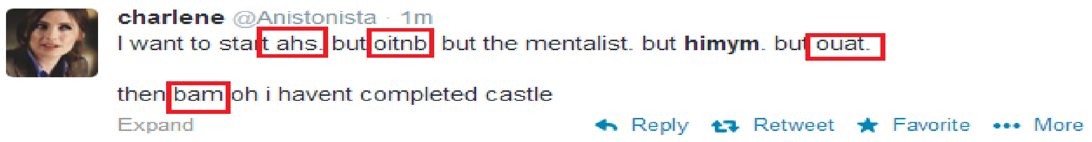
- Microblogging has become popular communication tool
- Opinion of the mass is important
  - Political party may want to know whether people support their program or not
  - A company might want find out the reviews of its products

Challenges that we face while performing sentimental analysis

- Tweets are highly unstructured and also non-grammatical

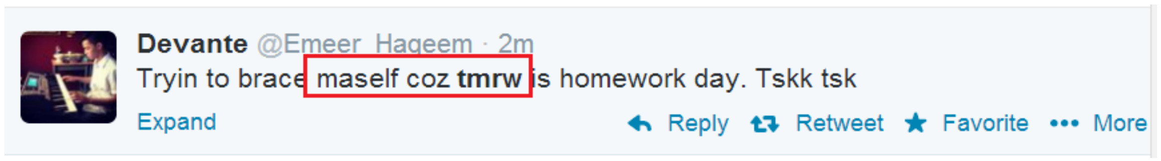

Abbey Flinn @abbeyflinn · Mar 24
MBP soaking up the sun ☀ pic.twitter.com/FwOh7JBGFU

- Out of Vocabulary Words


charlene @Anistonista · 1m
I want to start ahs. but oitnb. but the mentalist. but **himym**. but ouat.

then bam oh i havent completed castle

- Lexical Variation


Devante @Emeer_Haqeem · 2m
Tryin to brace maself coz **tmrw** is homework day. Tskk tsk

- Extensive usage of acronyms like asap, lol, afaik, emoji

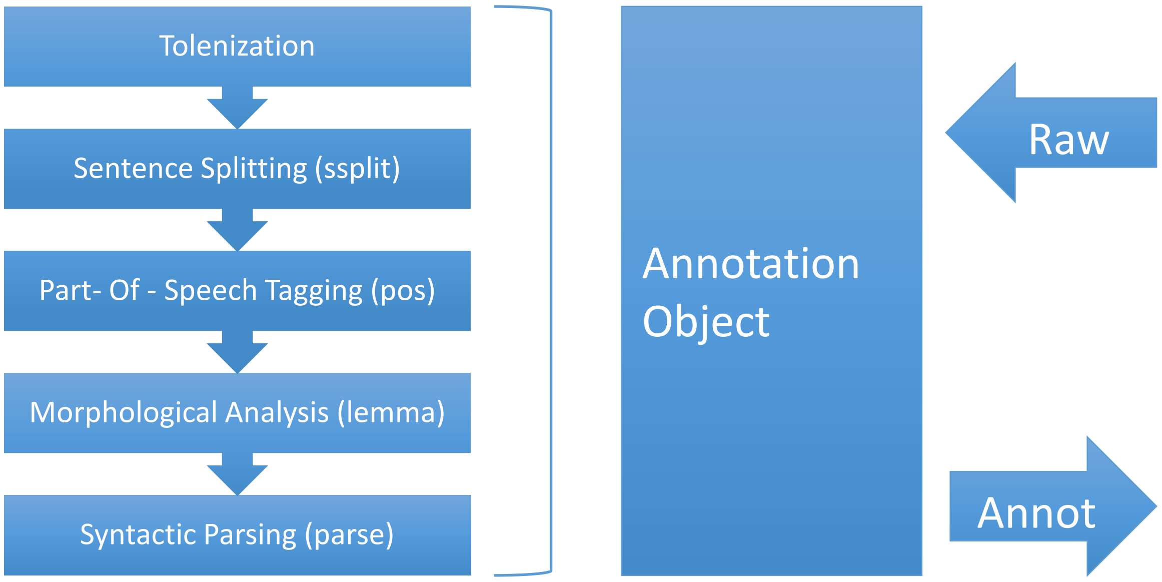
What is NLP?

- The attempt to use programming to read and understand the meaning of text
- The use of Natural Language processing, commonly denoted as NLP, with the intention of deriving "sentiment," or subjective information from text

Characterization of Tweets

- Binary Classification: It is a two way categorization i.e. positive or negative.
- 3-Tier: In this, Tweets are categorized as Positive, Negative and Neutral.
- 5-Tier: Tweets are bucketed in 5 Classes namely: Extremely Positive, Positive, Neutral, Negative and Extremely Neutral.

Approach – Stanford coreNLP


Tolenization
Sentence Splitting (ssplit)
Part- Of - Speech Tagging (pos)
Morphological Analysis (lemma)
Syntactic Parsing (parse)
Annotation Object
Raw
Annot

```
val props = new Properties()

props.setProperty("annotators", "tokenize, ssplit, pos, lemma, parse, sentiment")
```

Sentiment Function in Scala:

```
def detectSentiment(message: String): SENTIMENT_TYPE = {


  val pipeline = new StanfordCoreNLP(nlpProps)


  val annotation = pipeline.process(message)

  var sentiments: ListBuffer[Double] = ListBuffer()

  var sizes: ListBuffer[Int] = ListBuffer()


  var longest = 0

  var mainSentiment = 0
```

```
for (sentence <- annotation.get(classOf[CoreAnnotations.SentencesAnnotation])) {

  val tree = sentence.get(classOf[SentimentCoreAnnotations.AnnotatedTree])

  val sentiment = RNNCoreAnnotations.getPredictedClass(tree)

  val partText = sentence.toString


  if (partText.length() > longest) {

    mainSentiment = sentiment

    longest = partText.length()

  }


  sentiments += sentiment.toDouble

  sizes += partText.length


  println("debug: " + sentiment)

  println("size: " + partText.length)    }
```

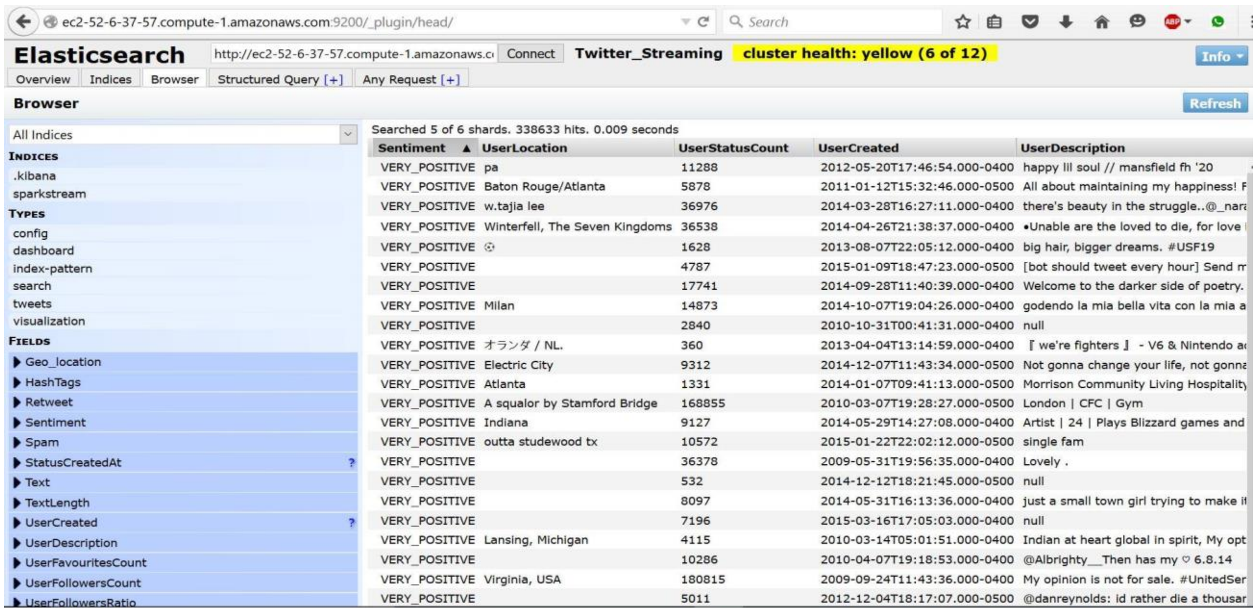
Detecting Sentiment:

- Collect the Tweet
- For each sentence in my tweet I calculate following:
    - Tree
    - Sentiment
    - Text size

Predicting Results

```
    • val averageSentiment:Double = {
if(sentiments.size > 0) sentiments.sum / sentiments.size

else -1

}

    • averageSentiment match {
  case s if s <= 0.0 => NOT_UNDERSTOOD

  case s if s < 1.0 => VERY_NEGATIVE
  case s if s < 1.5 => NEGATIVE

  case s if s < 2.5 => NEUTRAL

  case s if s < 3.3 => POSITIVE

  case s if s < 5.0 => VERY_POSITIVE

  case s if s > 5.0 => NOT_UNDERSTOOD
```

Storing Data into Elastic Search



ec2-52-6-37-57.compute-1.amazonaws.com:9200/_plugin/head/

Elasticsearch http://ec2-52-6-37-57.compute-1.amazonaws.c Connect Twitter_Streaming cluster health: yellow (6 of 12)

Browser

All Indices

Indices: .kibana, sparkstream

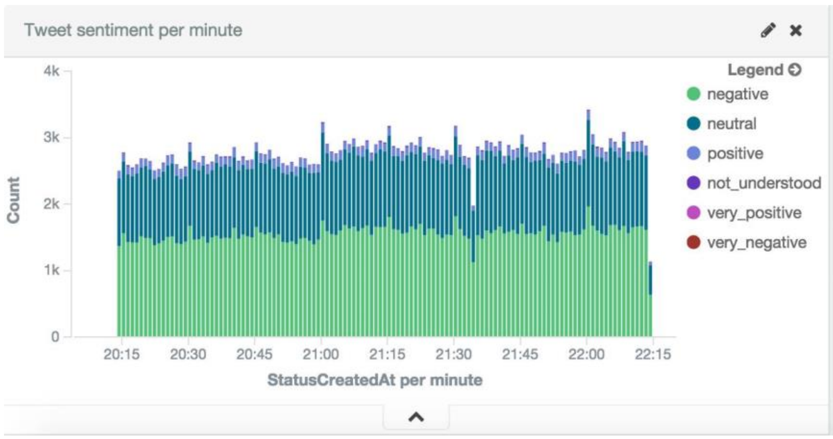
Types: config, dashboard, index-pattern, search, tweets, visualization

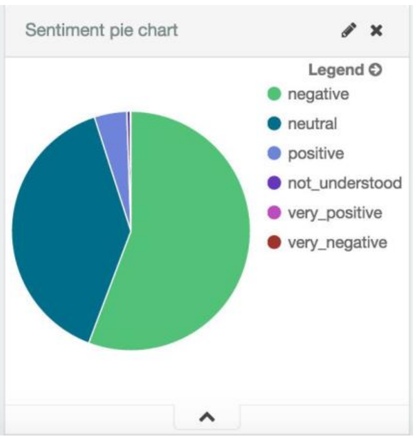
Fields: Geo_location, HashTags, Retweet, Sentiment, Spam, StatusCreatedAt, Text, TextLength, UserCreated, UserDescription, UserFavouritesCount, UserFollowersCount, UserFollowersRatio

Searched 5 of 6 shards. 338633 hits. 0.009 seconds

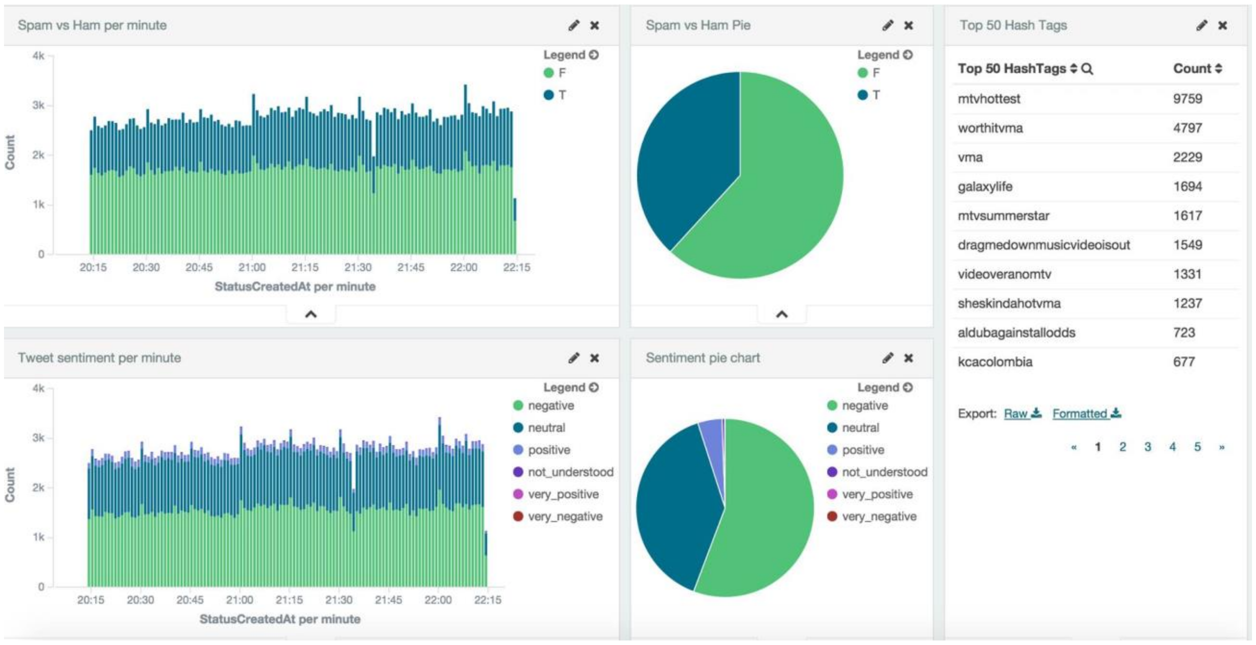
Sentiment ▲	UserLocation	UserStatusCount	UserCreated	UserDescription
VERY_POSITIVE	pa	11288	2012-05-20T17:46:54.000-0400	happy lil soul // mansfield fh '20
VERY_POSITIVE	Baton Rouge/Atlanta	5878	2011-01-12T15:32:46.000-0500	All about maintaining my happiness! F
VERY_POSITIVE	w.tajia lee	36976	2014-03-28T16:27:11.000-0400	there's beauty in the struggle..@_nar
VERY_POSITIVE	Winterfell, The Seven Kingdoms	36538	2014-04-26T21:38:37.000-0400	•Unable are the loved to die, for love
VERY_POSITIVE	⚽	1628	2013-08-07T22:05:12.000-0400	big hair, bigger dreams. #USF19
VERY_POSITIVE		4787	2015-01-09T18:47:23.000-0500	[bot should tweet every hour] Send m
VERY_POSITIVE		17741	2014-09-28T11:40:39.000-0400	Welcome to the darker side of poetry.
VERY_POSITIVE	Milan	14873	2014-10-07T19:04:26.000-0400	godendo la mia bella vita con la mia a
VERY_POSITIVE		2840	2010-10-31T00:41:31.000-0400	null
VERY_POSITIVE	オランダ / NL.	360	2013-04-04T13:14:59.000-0400	『 we're fighters 』 - V6 & Nintendo a
VERY_POSITIVE	Electric City	9312	2014-12-07T11:43:34.000-0500	Not gonna change your life, not gonn
VERY_POSITIVE	Atlanta	1331	2014-01-07T09:41:13.000-0500	Morrison Community Living Hospitality
VERY_POSITIVE	A squalor by Stamford Bridge	168855	2010-03-07T19:28:27.000-0500	London | CFC | Gym
VERY_POSITIVE	Indiana	9127	2014-05-29T14:27:08.000-0400	Artist | 24 | Plays Blizzard games and
VERY_POSITIVE	outta studewood tx	10572	2015-01-22T22:02:12.000-0500	single fam
VERY_POSITIVE		36378	2009-05-31T19:56:35.000-0400	Lovely .
VERY_POSITIVE		532	2014-12-12T18:21:45.000-0500	null
VERY_POSITIVE		8097	2014-05-31T16:13:36.000-0400	just a small town girl trying to make i
VERY_POSITIVE		7196	2015-03-16T17:05:03.000-0400	null
VERY_POSITIVE	Lansing, Michigan	4115	2010-03-14T05:01:51.000-0400	Indian at heart global in spirit, My opt
VERY_POSITIVE		10286	2010-04-07T19:18:53.000-0400	@Albrighty__Then has my ♡ 6.8.14
VERY_POSITIVE	Virginia, USA	180815	2009-09-24T11:43:36.000-0400	My opinion is not for sale. #UnitedSer
VERY_POSITIVE		5011	2012-12-04T18:17:07.000-0500	@danreynolds: id rather die a thousar

Twitter Sentiment per minute – Kibana Representation





Kibana Dashboard



Top 50 HashTags	Count
mtvhottest	9759
worthitvma	4797
vma	2229
galaxylife	1694
mtvsummerstar	1617
dragmedownmusicvideoisout	1549
videoveranomtv	1331
sheskindahotvma	1237
aldubagainstallodds	723
kcacolombia	677

Conclusion

- Spark streaming is a pretty powerful technology.
- Reusable code of batch and stream processing.
- Might be able to completely replace Hadoop MapReduce jobs Scala is a pretty powerful language.
- Even a complex application can be compile in less than 150 lines of code.
- We have combined the real stream data with our spam detection function for classification of tweets as Spam/Ham.
- We have combined real stream data with NLP dataset to get sentiment analysis

Future Work:

Join the data stream with Mlib library to enhance Spam detection

References

- Evaluation of the Effect of Spam on Twitter
  http://homepages.gac.edu/~lyu/Grant_paper.pdf
- Detecting malicious tweets in trending topics
  http://nlp.uned.es/~lurdes/araujo/eswa13_malicious_tweets.pdf
- http://mashable.com/2013/11/08/twitter-spambots/
- http://dev.datasift.com/docs/csdl/csdl-examples/filtering-twitter-spam
  http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf
- http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf