CKME136 Capstone Project

A big Data application for Real Time classification of Symptoms of illness in Toronto, using twitter data

by Leotis Buchanan(LeotisBuchanan@gmail.com)

June 17, 2015

Problem

1. Quickly Detecting disease outbreaks using user social media posts.
2. Handling and processing large quantity of data in real time(volume, velocity).
3. Using the data to predict future illness outbreak.

Dataset and datasource

1. Twitter via their data stream api.

The schema for a tweet

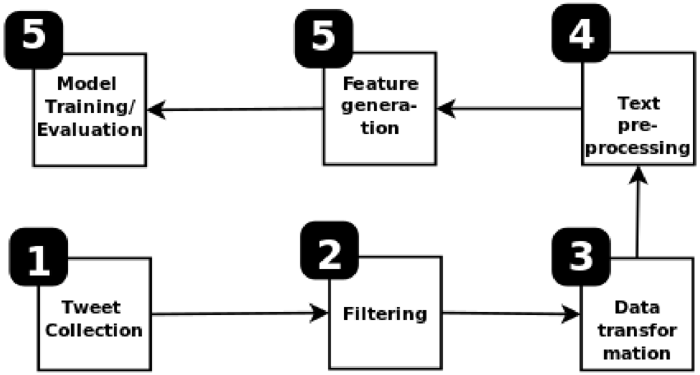
The schema/structure of the tweet data collected was generated and printed using the following snippet of code:

```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
# Create the DataFrame
df = sqlContext.jsonFile("twitter_stream.20150519-084907.json")
# Show the content of the DataFrame
df.show()
# Print the schema in a tree format
df.printSchema()
```

The data format of a single tweet.

```
root
 |-- _corrupt_record: string (nullable = true)
 |-- contributors: string (nullable = true)
 |-- coordinates: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- created_at: string (nullable = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
```

Approach



1. Tweet collection
2. Tweet Cleaning and transformation
3. Filtering
4. Data transformation
5. Text Preprocessing and feature generation

Training the Classifier

1. Generate feature vectors each tweet text.

```
def generatedHashedFeatures(tweet):
   htf = HashingTF()
   lp = LabeledPoint(tweet.label, htf.transform(tweet.text))
   return lp
```

2. Manually labelled about 1000 tweets.

> *user_id_100,,,en,Tue May 26 00:58:25 +0000 2015, MamaJaws not everyone I m itching to formulate an opinion but cannot because there isn t enough data ,0*

> *user_id_01,,,en,Thu May 28 16:59:32 +0000 2015, Another dialysis stay under my belt lots of itching and cramping Lord have mercy ,1*

1. Split data in training and test data.

   ```
   training, test = data.randomSplit([0.6, 0.4], seed = 0)
   ```

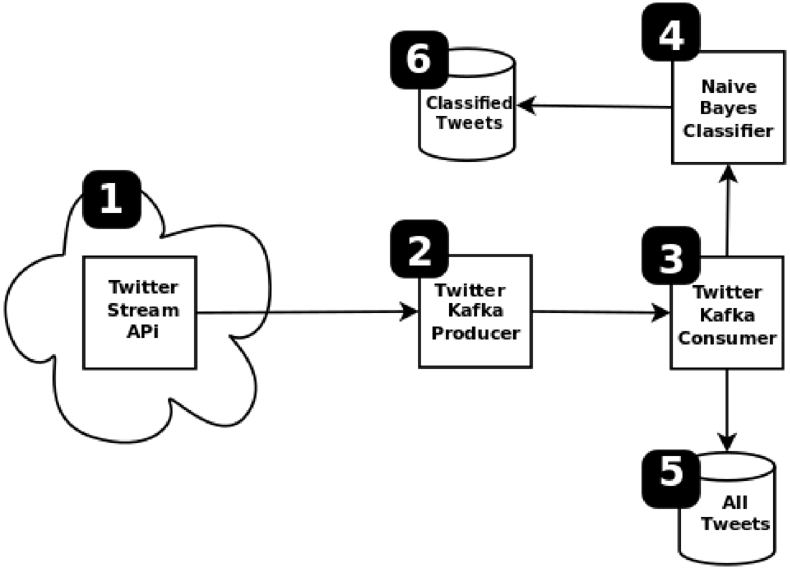
2. Train and persist a naive bayes model.

   ```
   model = NaiveBayes.train(training, 1.0)
   ```

Classifying tweets in Real Time

1. **Apache Spark** - Apache Spark is a fast and general-purpose cluster computing system
2. **mlib** - Machine Learning Library
3. **kafka**- A high-throughput distributed messaging system.
4. **twitter**- streaming api.
5. **deployed to AWS**

Application Architecture



Conclusion and results

1. Created a spark application that streams and classify tweets in real time.
2. Trained naive bayes classifier model.
3. I have made the source code for the project available on my github repo.

Future work

1. Create android app to visualize the output of the application.
2. Deploy the application to databricks cloud.
3. Incooperate other datasources.
4. Fix bugs etc.

Questions ?