Deep Text Stream Analysis

Harilal Orunkara Poyil < harilal@kth.se>
Gopal Arun Tathe <tathe@kth.se>

Project Report

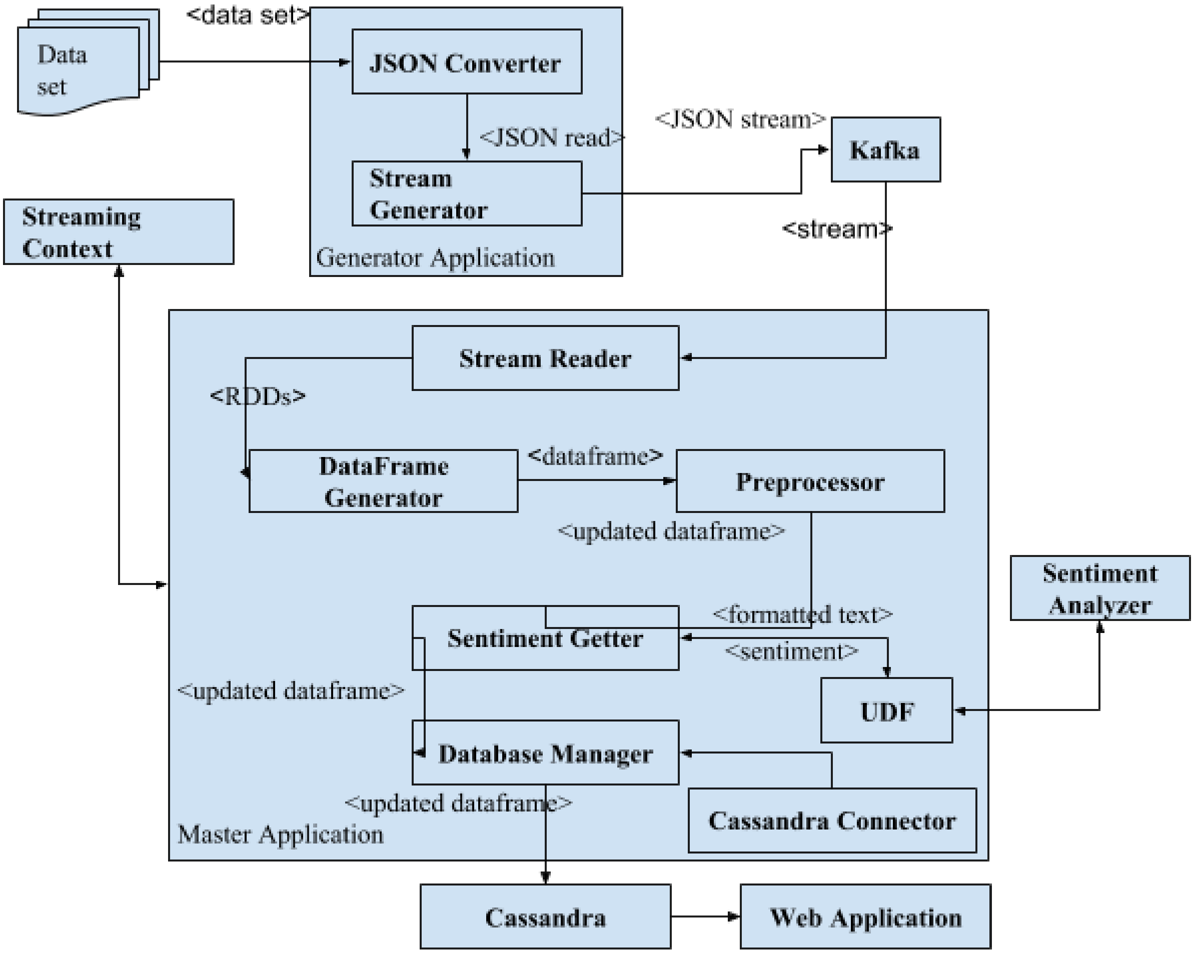
Introduction

Extracting valid, novel, useful/actionable, understandable information from large amount of data is always having significance in various domains. The data can come from various sources, can have various forms, structured or unstructured and can be either static or stream. Collecting, storing, pre-processing, analyzing and communicating the results bring lot of challenges. It's observed that the methods used at each stages of processing vary based on the behaviour of data.

In this project we analyze stream of text data. We estimate the sentiment of each text and store in a database. We also provide a visualization for sentiment distribution.

System Architecture

The architecture of system is given in Figure(i)



Figure(i). System architecture

The system contains following main components.

i) Stream generator
ii) Message broker
iii) Master application
iv) Sentiment analyzer
v) Data storage

The stream generator reads text data from dataset[1] and create a JSON string in the form of following sample format.

```
{
        id = "1467815924",
        date = "Mon Apr 06 22:19:49 PDT 2009",
        text = "@alielayus I want to go to promote GEAR AND GROOVE but unfortunately no ride there..."
}
```

Then this JSON will send to the message broker - Kafka. The master application receives this message and convert the underlying RDDs to Dataframes. As part of preprocessing we remove URLs from text as it's seldom contribute to analysis. In this step a new column named *formatted_text* is added to the dataframe which is then passed to sentiment analysis stage. For sentiment estimation, we are using Stanford NLP[2] which use Recursive Neural Tensor Networks and the Sentiment Treebank. Using user defined function[UDF], we integrate the sentiment analysis to this spark streaming application. The sentiment is get added to the dataframe and get pushed to Cassandra database.

For analytics purpose, Python notebook is used. In which basic analysis is shown. It reads data stored from Cassandra database and transform it into pandas data frame to carryout operation.

Tools

Big Data: Apache Spark, Cassandra, Kafka
Visualization tools: D3 Visualization, Tableau visualization.
Development tools: Scala, Python, and JavaScript
Natural Language Processing Algorithms : Stanford NLP

Data Set

Source: **Sentiment140 dataset with 1.6 million tweets** [3]

Running the application

Source code : https://github.com/HarilalOP/DeepTextStreamAnalyzer.git

Kafka

```
//export environment variables
export KAFKA_HOME="/usr/local/kafka"
export PATH=$KAFKA_HOME/bin:$PATH

//start zookeeper
$KAFKA_HOME/bin/zookeeper-server-start.sh $KAFKA_HOME/config/zookeeper.properties
```

//start kafka server

$KAFKA_HOME/bin/kafka-server-start.sh $KAFKA_HOME/config/server.properties

//create Kafka topic

$KAFKA_HOME/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic text_anlyz

Cassandra

//export environment variables

export CASSANDRA_HOME="/usr/local/cassandra"
export PYTHONPATH="/home/harilal/anaconda2/bin/python"
export PATH=$PYTHONPATH/bin:$CASSANDRA_HOME/bin:$PATH

//Start Cassandra in the foreground

$CASSANDRA_HOME/bin/cassandra -f

//Start the cqlsh prompt

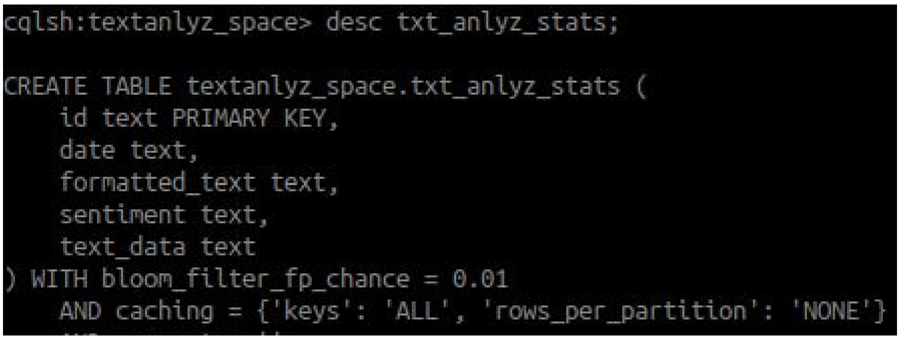
$CASSANDRA_HOME/bin/cqlsh

//Create keyspace(optional as it's getting created from the code)

create keyspace textanlyz_space with replication = {‘class’: ‘SimpleStrategy’, ‘replication_factor’: 1};

//Create table(optional as it's getting created from the code)

use textanlyz_space;
CREATE TABLE IF NOT EXISTS textanlyz_space.txt_anlyz_stats (id text PRIMARY KEY, date text, text_data text, formatted_text text, sentiment text);
desc txt_anlyz_stats; //figure(i)



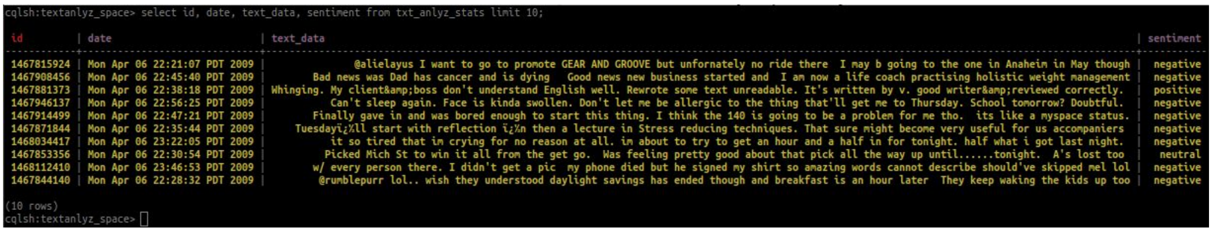
```
cqlsh:textanlyz_space> desc txt_anlyz_stats;

CREATE TABLE textanlyz_space.txt_anlyz_stats (
    id text PRIMARY KEY,
    date text,
    formatted_text text,
    sentiment text,
    text_data text
) WITH bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
```

Figure(i). Table description

//Check table content

select * from txt_anlyz_stats;
select id, date, text_data, sentiment from txt_anlyz_stats limit 10; //figure(ii)



```
cqlsh:textanlyz_space> select id, date, text_data, sentiment from txt_anlyz_stats limit 10;

 id         | date                         | text_data                                                                                                                                                   | sentiment
------------+------------------------------+-------------------------------------------------------------------------------------------------------------------------------------------------------------+-----------
 1467815924 | Mon Apr 06 22:21:07 PDT 2009 |              @alielayus I want to go to promote GEAR AND GROOVE but unfornately no ride there  I may b going to the one in Anaheim in May though |  negative
 1467908456 | Mon Apr 06 22:45:40 PDT 2009 |       Bad news was Dad has cancer and is dying   Good news new business started and  I am now a life coach practising holistic weight management |  negative
 1467881373 | Mon Apr 06 22:38:18 PDT 2009 | Whinging. My client&amp;boss don't understand English well. Rewrote some text unreadable. It's written by v. good writer&amp;reviewed correctly.  |  positive
 1467946137 | Mon Apr 06 22:56:25 PDT 2009 |          Can't sleep again. Face is kinda swollen. Don't let me be allergic to the thing that'll get me to Thursday. School tomorrow? Doubtful.  |  negative
 1467914499 | Mon Apr 06 22:47:21 PDT 2009 |       Finally gave in and was bored enough to start this thing. I think the 140 is going to be a problem for me tho.  its like a myspace status. |  negative
 1467871844 | Mon Apr 06 22:35:44 PDT 2009 |    Tuesdayï¿½ll start with reflection ï¿½n then a lecture in Stress reducing techniques. That sure might become very useful for us accompaniers  |  negative
 1468034417 | Mon Apr 06 23:22:05 PDT 2009 |          it so tired that im crying for no reason at all. im about to try to get an hour and a half in for tonight. half what i got last night.  |  negative
 1467853356 | Mon Apr 06 22:30:54 PDT 2009 |         Picked Mich St to win it all from the get go.  Was feeling pretty good about that pick all the way up until......tonight.  A's lost too  |   neutral
 1468112410 | Mon Apr 06 23:46:53 PDT 2009 |       w/ every person there. I didn't get a pic  my phone died but he signed my shirt so amazing words cannot describe should've skipped mel lol |  negative
 1467844140 | Mon Apr 06 22:28:32 PDT 2009 |        @rumblepurr lol.. wish they understood daylight savings has ended though and breakfast is an hour later  They keep waking the kids up too |  negative

(10 rows)
cqlsh:textanlyz_space> 
```

Figure(ii). Table content

Stream

//Generate streaming input

```
cd Project/DeepTextStreamAnalyzer/src/main/generator
sbt run
```

Application

//Run the application

```
cd Project/DeepTextStreamAnalyzer/src/main/analyzer
sbt run
```

Analytics Notebook

Install Cassandra Python driver before running notebook commands.

References

[1].“Sentiment140’ - A Twitter Sentiment Analysis Tool, Sentiment140, help.sentiment140.com/for-students.

[2]. Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

[3]. Go, A., Bhayani, R. and Huang, L., 2009. “Twitter sentiment classification using distant supervision. CS224N Project Report”, Stanford, 1(2009), p.12.