**Ambiguity Reduction and Sentimental Analysis**

**Date 04/24/2017**

**Team Members:**

**Abhijit Shinde**

**Reema D’mello**

**Tushar Goel**

Contents

[Introduction 3](#_Toc480783640)

[Context (or scope) 3](#_Toc480783641)

[Objective 4](#_Toc480783642)

[Technologies: 4](#_Toc480783643)

[AWS Platform Installation( AWS/ Hadoop / Flume / Hive /Mahout / Odbc drivers for Tableau): 4](#_Toc480783644)

[Hive Installation: 7](#_Toc480783645)

[Flume Installation: 8](#_Toc480783646)

[Dataset & Data cleansing, Extraction of live data. 8](#_Toc480783647)

[Live data stream: 8](#_Toc480783648)

[Mahout Classification (Naive Bayes model). 9](#_Toc480783649)

[NLTK Sentimental analysis using python: 11](#_Toc480783650)

[Project workflow and Hive table setup. 12](#_Toc480783651)

[Reporting and dashboard for data visualization: 16](#_Toc480783652)

[References: 16](#_Toc480783653)

# Introduction

This project report is intended to document the final project Ambiguity Reduction and Sentimental Analysis for the class Engineering Big Data Systems. The project was thought through consultation with Professor Dino Konstantopoulos and Teaching Assistants Vasanti Mahajan and Paritosh Arora.

## Context (or scope)

The volume, velocity, and variety of data generated on the internet have made it necessary for us to mine internet and extract required amount of data for our business to utilize it effectively. Big Data pioneers to develop the empowering technology. The amount and variety of data generated on the internet in one second is tremendous and is only bound to increase. While most businesses today generate the huge volume of data, it does mean that Big Data application is applicable for them. With the inception of [Internet of Things](https://en.wikipedia.org/wiki/Internet_of_Things), and [Industry 4.0](https://en.wikipedia.org/wiki/Industry_4.0) the prospects to acquire previously uncaptured data is immense.

The project aims at creating end to end big data application to mine required data by reducing ambiguity of data in cases where the name of the show also resembles some other entity. As for the scope of the project the primary user are TV producers and all beneficiary who care about entity where the name of the show also resembles some other entity. Sentimental analysis on the data found can go incorrect due to ambiguous social media data which has same names. In this case we are using Twitter’s platform and Facebook data for classifying and analyzing user’s sentiments using big data technologies for a show.

Twitter and facebook are the Social networking and micro blogging service. It enables users to send and read messages where messages known as “status”, “comments” and "tweets". The data is continuously streaming online and may contain rich information about people’s preferences. People share their thoughts about TV shows.

# Objective

Objective of the project is to stream live data from social media save it on hadoop infrastructure, classify using mahout classifier in parallel and perform sentimental analysis for TV series Silicon Valley (in this use case), future we are creating a tableau report as visualization.

# Technologies:

* Amazon web services
* Apache Hadoop
* Apache Flume
* Apache Flume
* Apache Hive
* Python 2.7 with NLTK package
* Tableau
* Unix scripting

# AWS Platform Installation (AWS/ Hadoop / Flume / Hive/Mahout / Odbc drivers for Tableau):

**Part 1.**

**Configuring Hadoop in Fully distributed mode:**

Fully distributed mode allows us to utilize all features of Hadoop, I am using Amazon EC2 to configure Hadoop cluster.

Following are the component of Hadoop cluster ( 5 nodes):

* NameNode (Master)/ SecondaryNameNode
* DataNode (Slave1)
* DataNode (Slave2)
* DataNode (Slave3)
* DataNode (Slave4)

Tools used:

* Amazon EC2
* Putty
* Putty Keygen
* Winscp

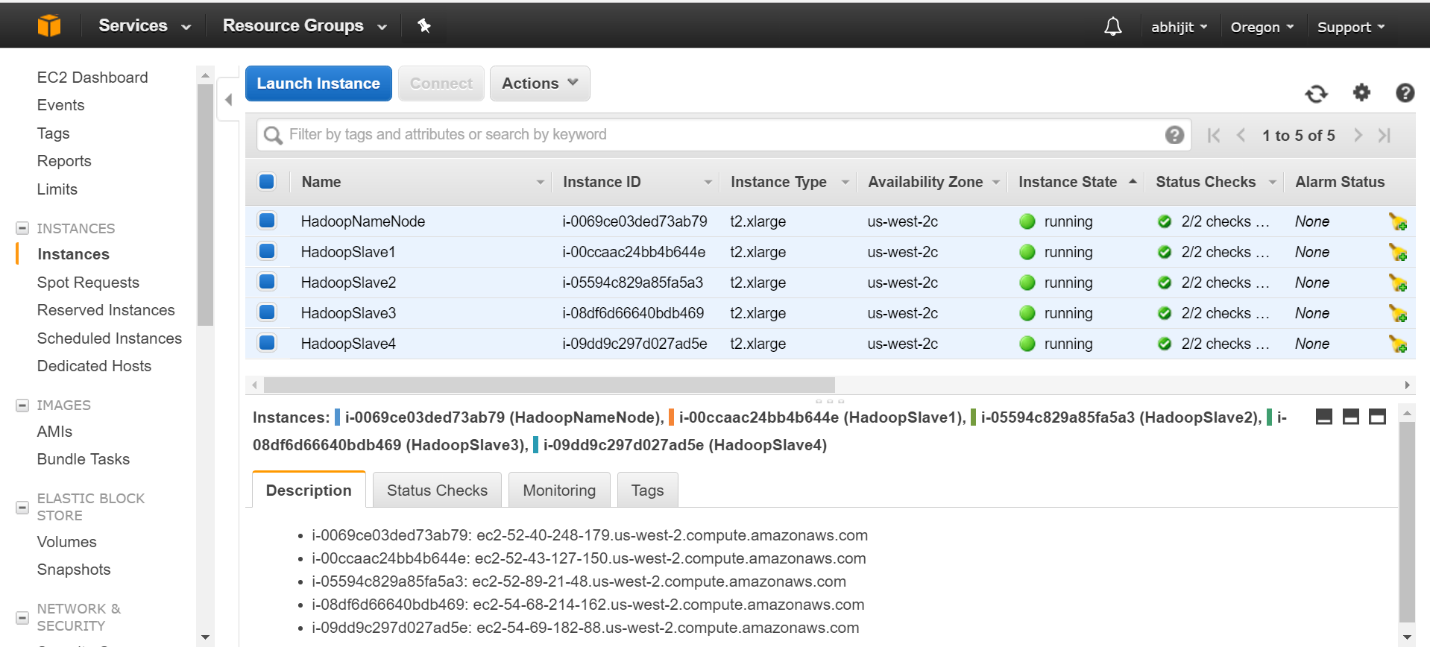
Set up of Hadoop cluster on Amazon EC2 AWS/ Setting up instances:

Selected 5 instance, t2Xlarge which has 8GB Ram and Dual core process with 150GB SSD Space each. Initial smaller instances like t2 micro with 1GB Ram did and 30 GB SSD did not work. So I terminated these instances.

Assigned dedicated IP address as that configuration will not be lost if we stop aws instance. For this I used Elastic IP.

Managed and generated public and private keys using puttykey gen and pem file.

Created Security groups and launched instance.

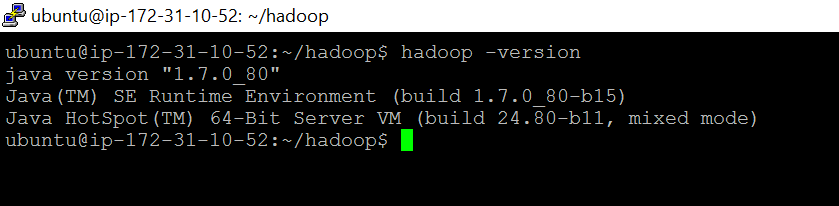


After creating instance connected each instance using putty and configured its host and assigned hostname as per amazon DNS instance.

Updated all packages on instance.

Downloaded and installed Java

Downloaded and Installed Hadoop 1.2.1



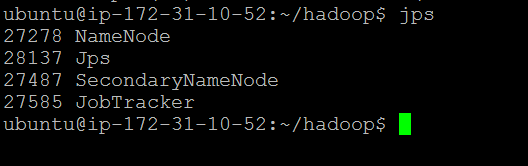
For communication purpose between different data nodes and master nodes copied pem file( keys)

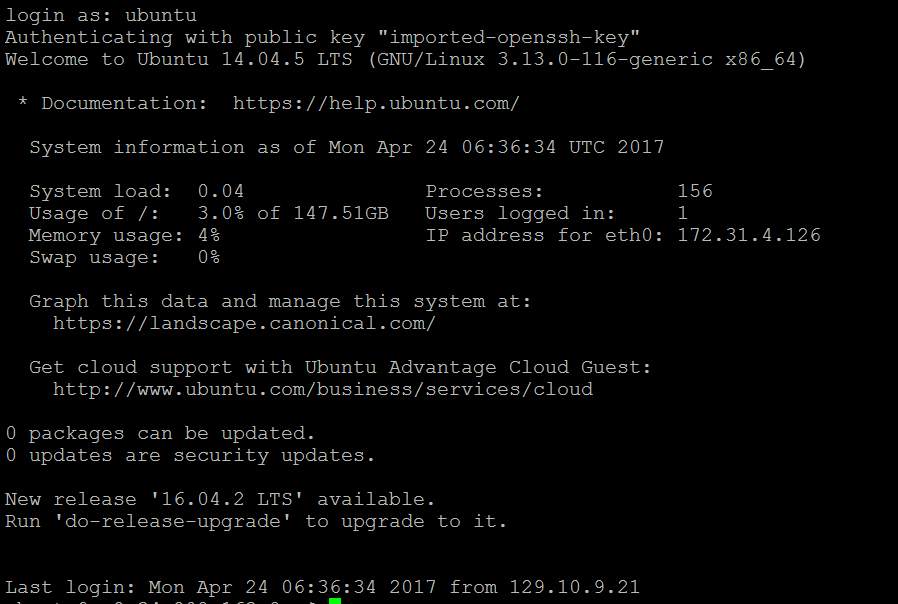
Configured Environment variables like java home and HADOOP\_CONF

Configured password less SSH so that nodes can communicate with each other.

Updated Hadoop configuration files like **hadoop-env.sh, core-site.xml, hdfs-site.xml, mapred-site.xml,** **master** file and **Slaves** file on each node.

Formatted name node and Started all instance using start-all.sh





## Hive Installation:

1. Hive: https://www.edureka.co/blog/apache-hive-installation-on-ubuntu

http://doctuts.readthedocs.io/en/latest/hive.html

2. wget http://apache.mirrors.pair.com/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz

export HIVE\_HOME="/home/ubuntu/hive/hive"

export PATH=$HIVE\_HOME/bin:$PATH

export HADOOP\_HEAPSIZE=1000

export HIVE\_CONF\_DIR="/home/ubuntu/hive/hive/conf"

## Flume Installation:

1. -- flume reference http://hadooptutorial.info/apache-flume-installation/

2. wget http://archive.apache.org/dist/flume/1.5.0/apache-flume-1.5.0-bin.tar.gz

tar -xzf apache-flume-1.5.0-bin.tar.gz

export FLUME\_HOME="/home/ubuntu/flume/flume"

export FLUME\_CONF="$FLUME\_HOME/conf"

export FLUME\_CLASSPATH="$FLUME\_CONF"

export PATH="$FLUME\_HOME/bin:$PATH"

# Dataset & Data cleansing, Extraction of live data.

(Twitter and Facebook data scraping using python and Flume and data cleansing with unix script)

Most of Social media group today have APIs available for data import, We have used Facebook and Twitter data using APIs and Python script. Following are the library in python we used, tweepy, Json, URL lib 2

Data that we used to training and testing initial models is from 2013, around 85000 Facebook comments/Status were scraped and 15000 user tweets were scraped by twitter API.

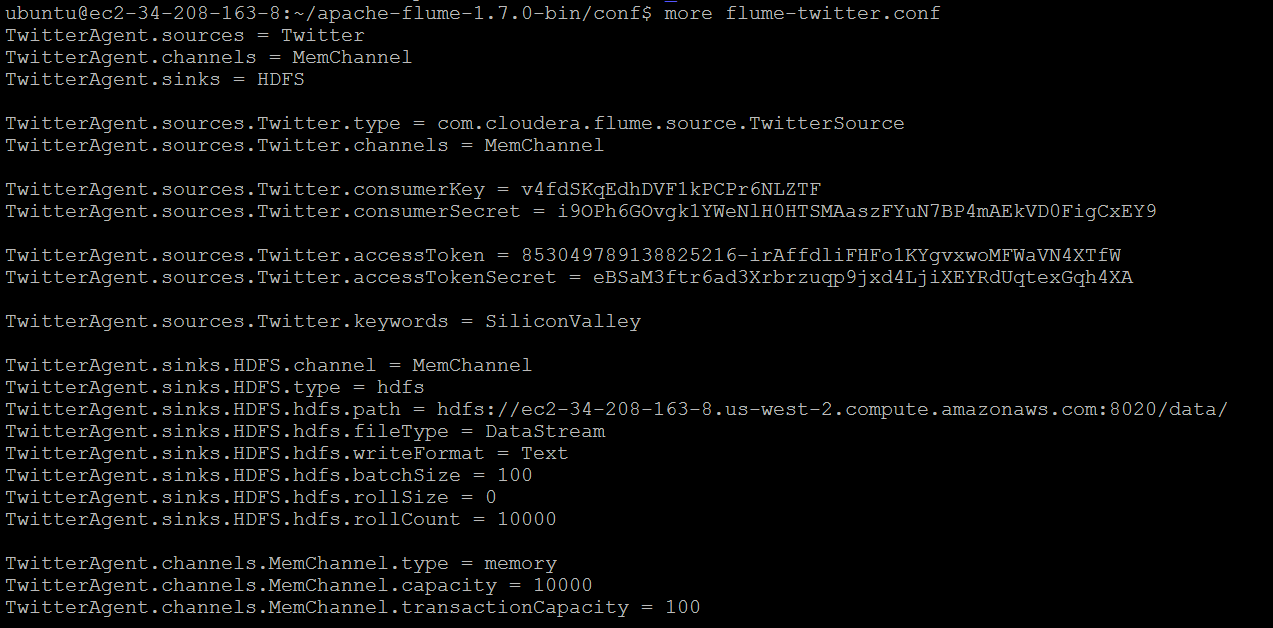
Data cleansing was performed initially, the source of out data was from authorized twitter Handle like SiliconValHBO, SiliconHBO, SiliconVallism. The character set was converted to UTF 8 using as Data was labeled for training.

Since we have data from multiple sources we have different sets of features:

1. User\_ID, Screen\_name, created\_at, tweet\_type, text
2. User ID, created\_at, text
3. Comment\_id, status\_id, comment\_message, comment\_likes
4. Status\_id, status\_message, number\_of\_shares, status\_type, status\_link, permalink\_url, status\_published, num\_reactions, num\_comments, num\_shares, num\_likes, num\_loves, num\_wows, num\_hahas, num\_sads, num\_angrys

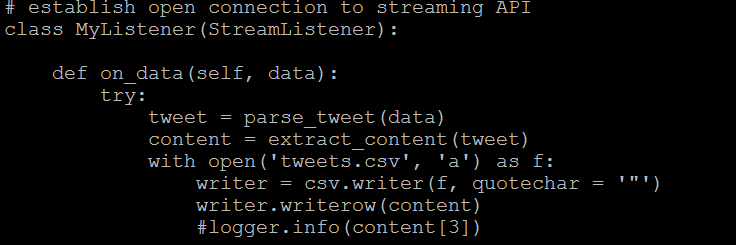
## Live data stream:

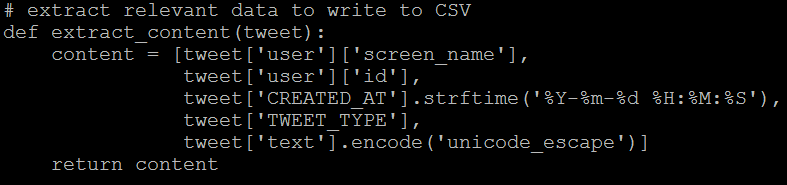
Configured Flume Agent for twitter:



Following are the **live data** scraping steps: (JSON Format)

1. Create developer account, download the keys for connection.
2. Create python scripts for fetching along with our credentials.
3. Most of data is in JSON format, which we parse to csv internally using parser script in python.
4. We have fetched the data based on multiple hashtags, page ID, user id, comments and group ID. (Only public profile data is available)
5. Data is stored in directory, fetching process and scheduling is managed by unix script.



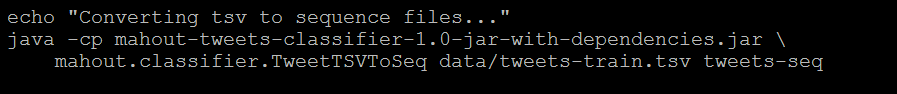


# Mahout Classification (Naive Bayes model).

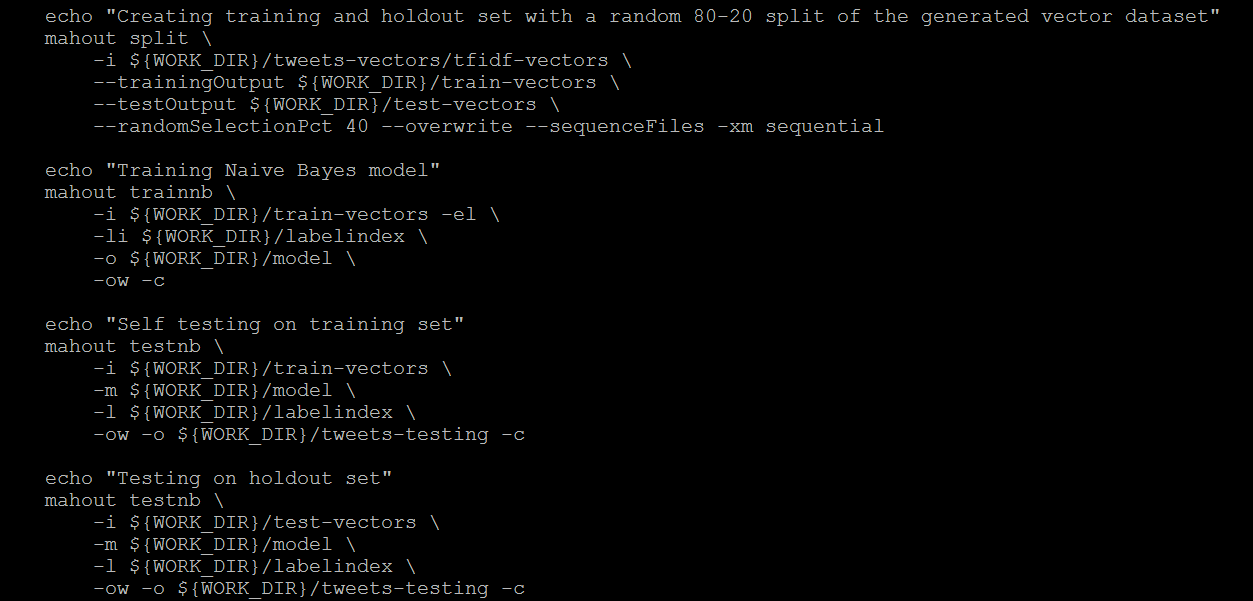
Using Mahout command and Hadoop directory structure we performing.

Following are the steps:

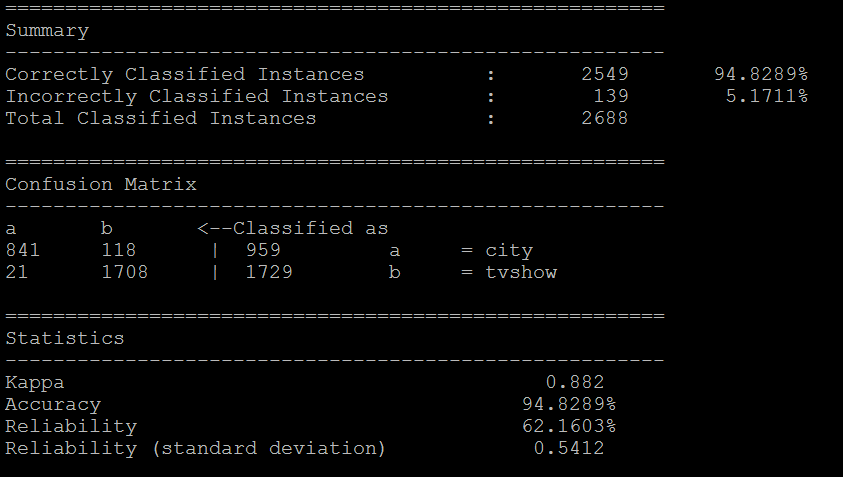
1. Converting files to sequence as mahout process files in sequence



1. Vectoring files converting word2vec using mahout seq2sparse
2. Performing training, dividing data into 80/20 (splitting, training, testing)



**Mahout model accuracy:**



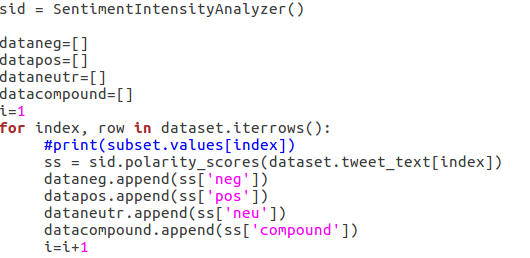
# NLTK Sentimental analysis using python:

NLTK is a leading platform for building Python programs to work with human language data. We have used the Vader sentiment analysis which computes not only the positive and negative words but also the intensity of those words and their hidden expressions. Weights are assigned to the words based on their sentiment intensity and this helps in forming a good analysis.

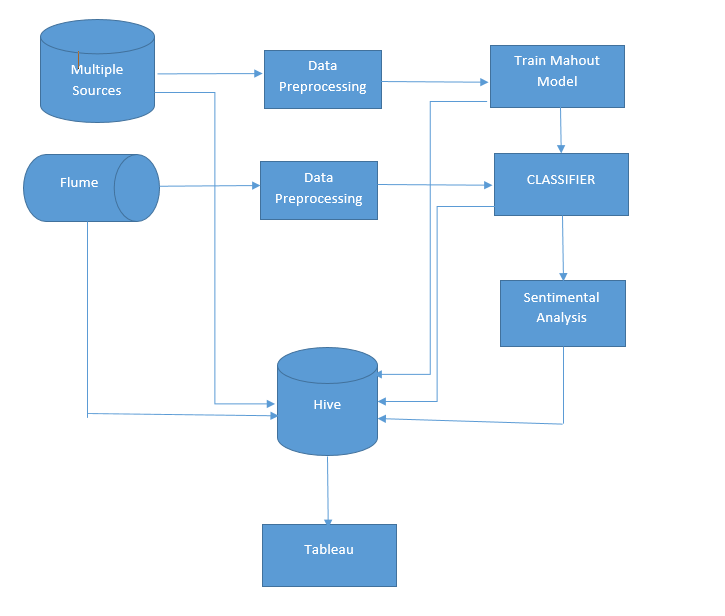
We have to download the nltk() package and install Vader Sentiment.

> pip install vaderSentiment

We get the polarity scores of each sentiment namely positive, negative and neutral for each sentence.



# Project workflow and Hive table setup.

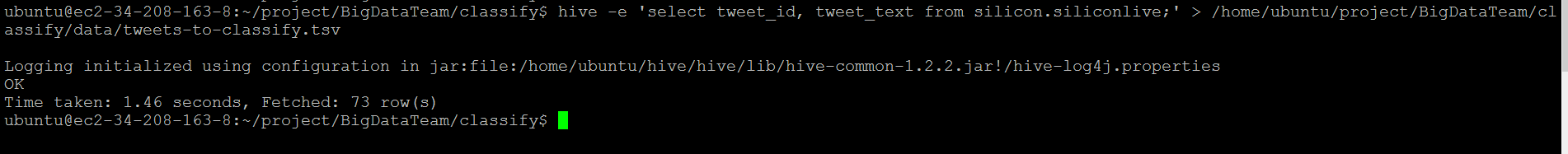


Multiple Hive tables are created and to save all data before and after analysis is stored.

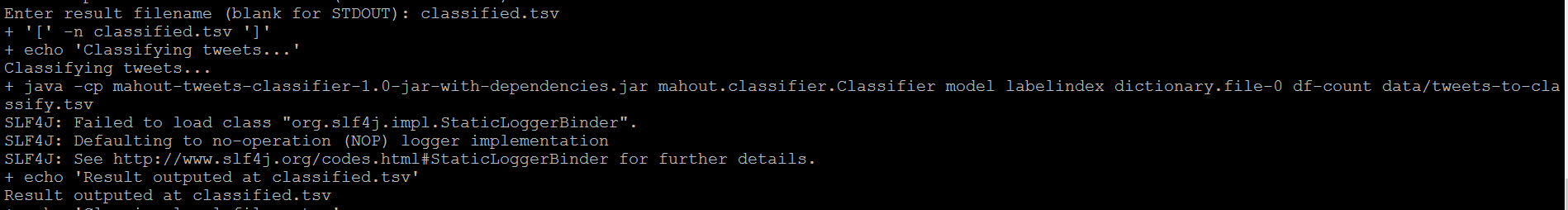


Unix script run.sh performs all the operations automatically: Below are the sequence of steps and output:

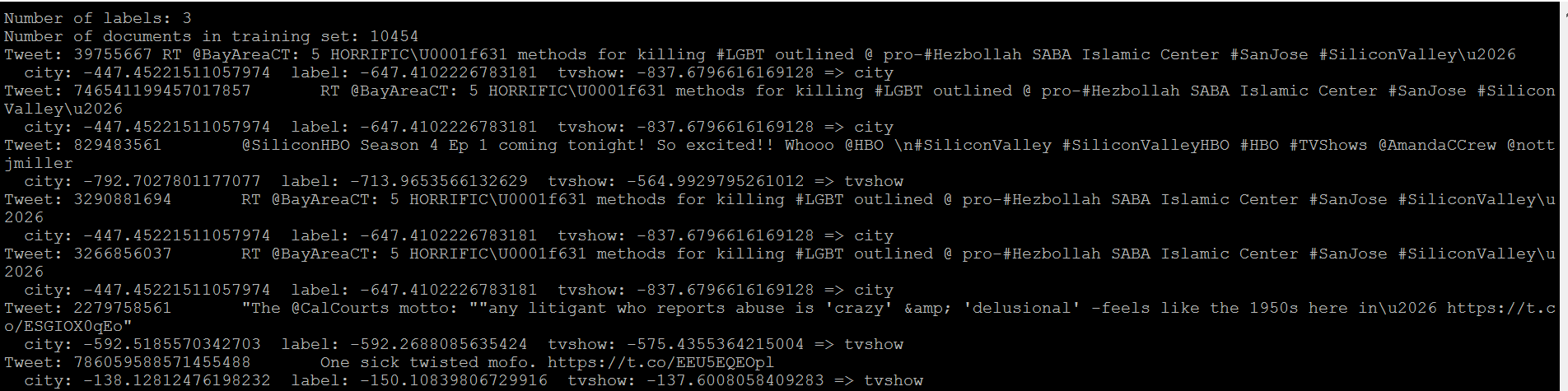
1. Fetching new data: in this example 73 rows



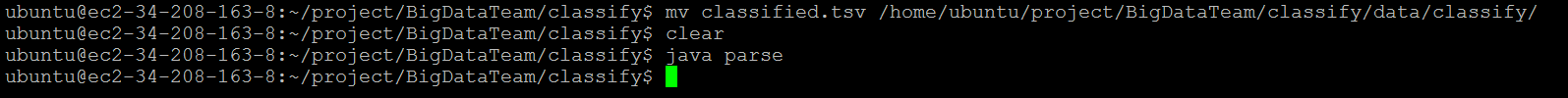
1. Classifying data using mahout classifier jar:



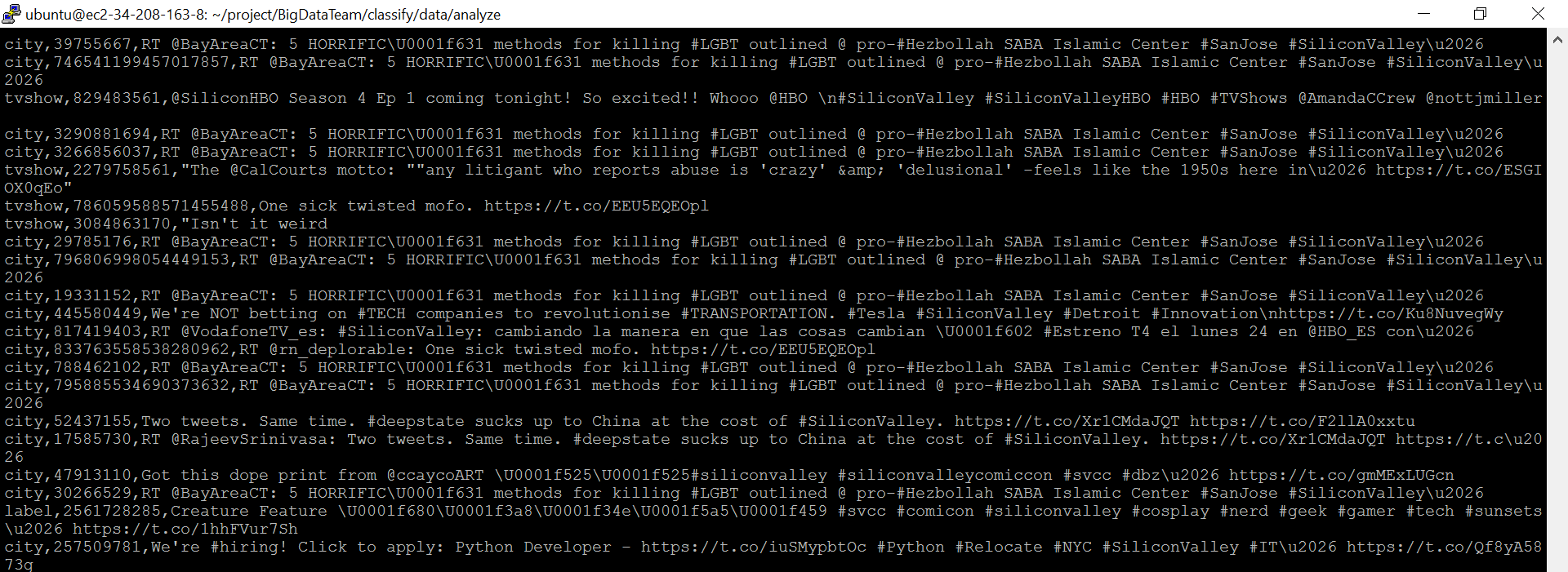
1. Output of mahout classifier provides weight to city and TV show in text format:



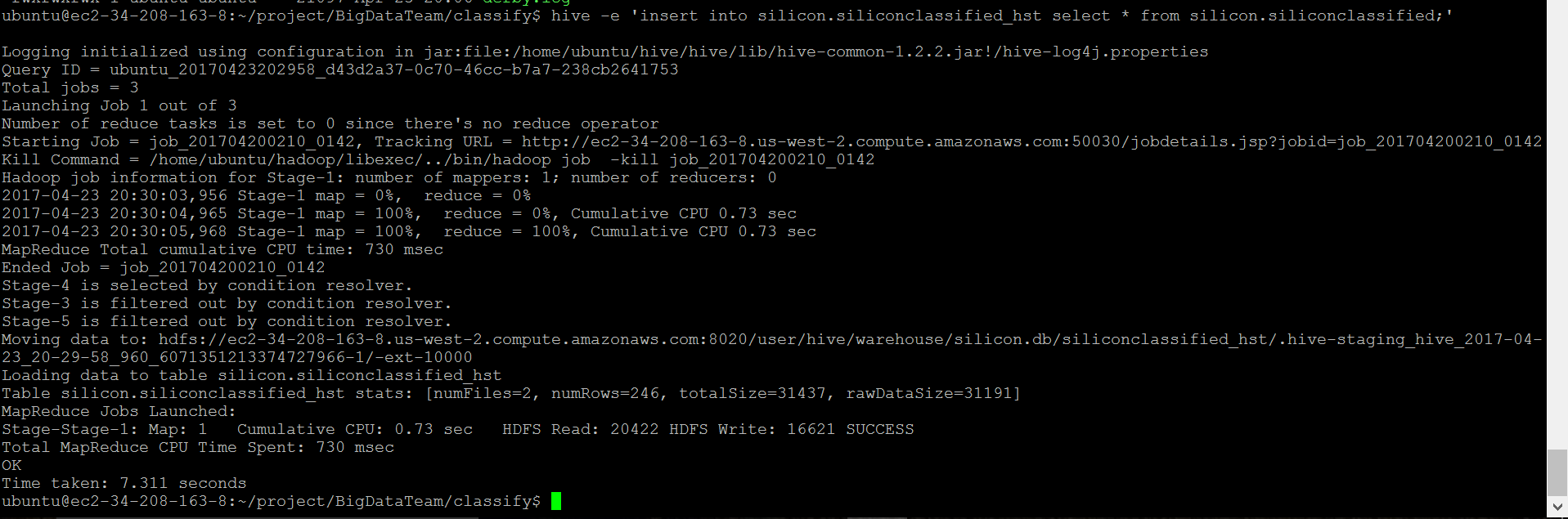
1. Parsing classified data: using java 🡪 output classified data so that format is corrected and data is in relational format:



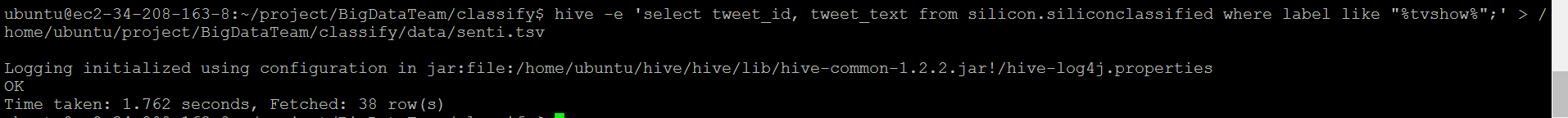
1. Parsed data sample in relational format in tsv:



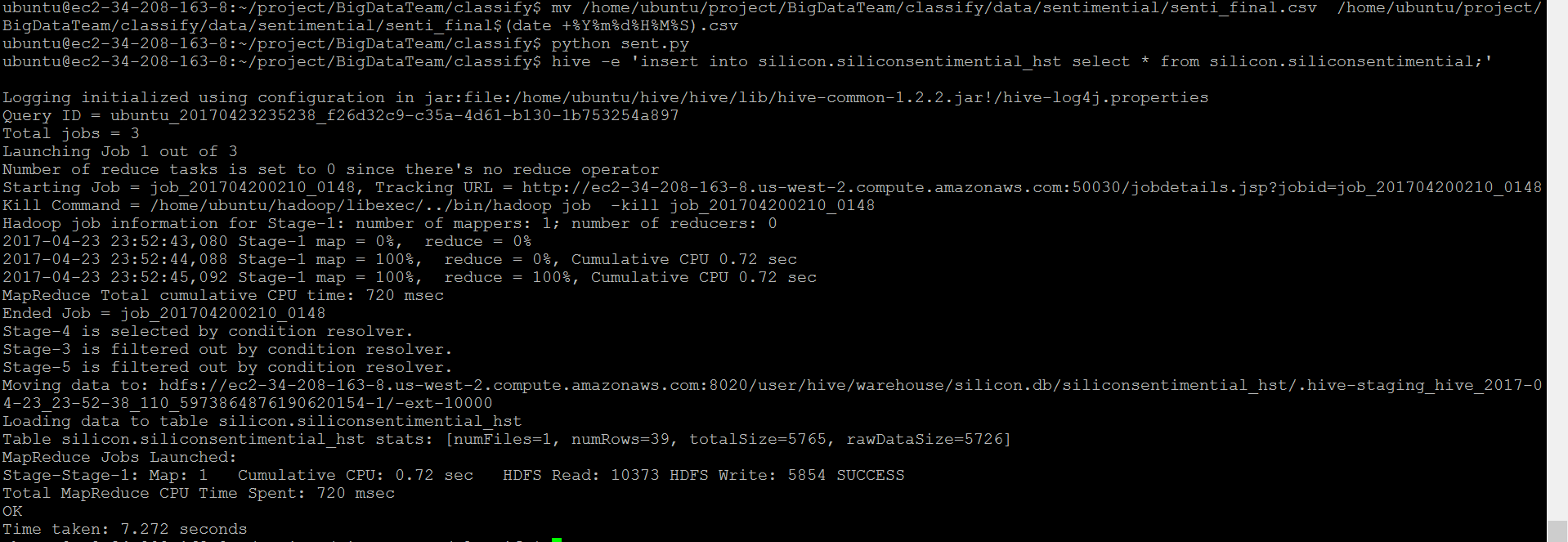
1. Inserting classified data (tabular format in hive tables)



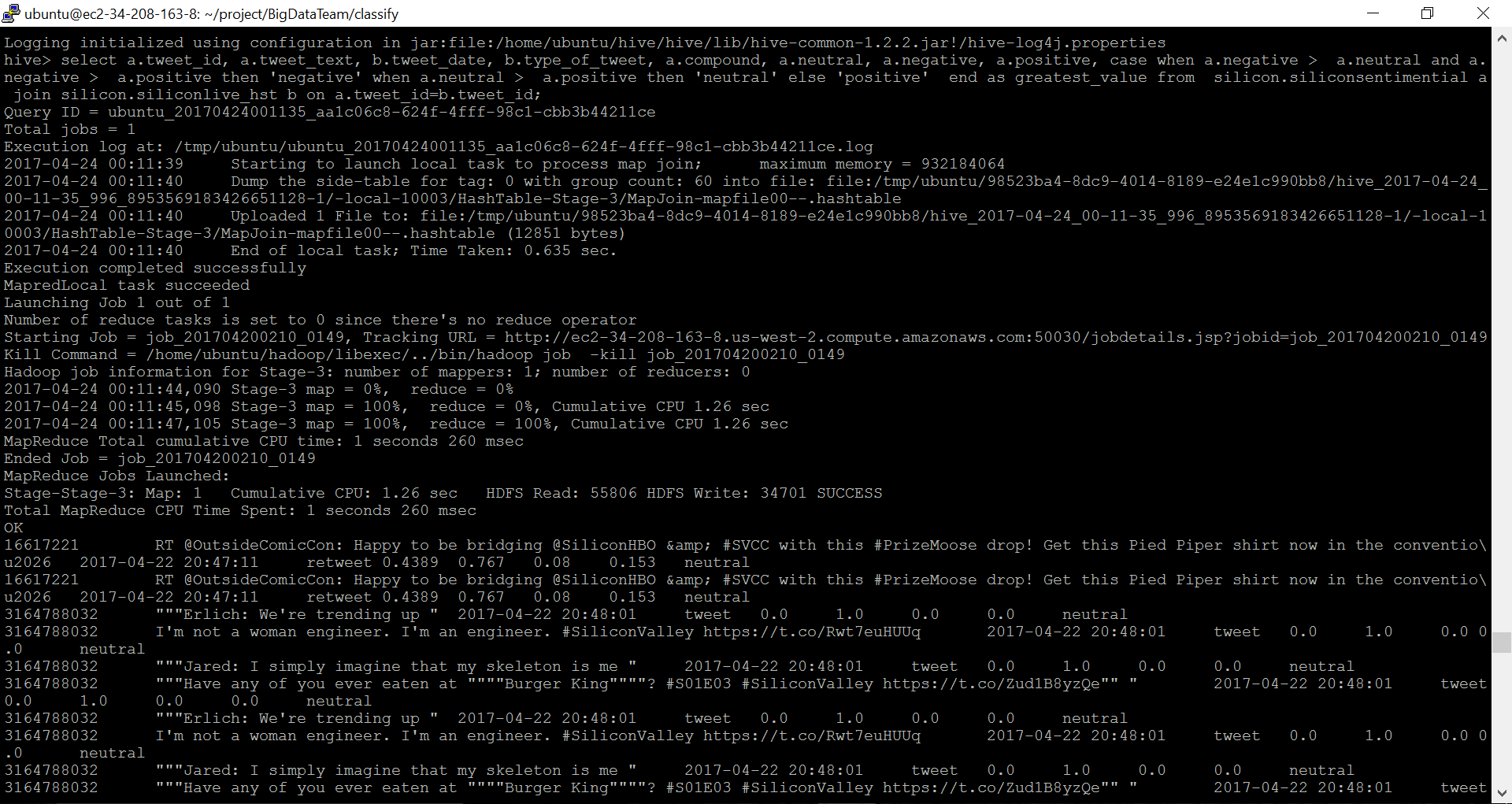
1. Using hive table exporting data for sentimental analysis: (38 records)



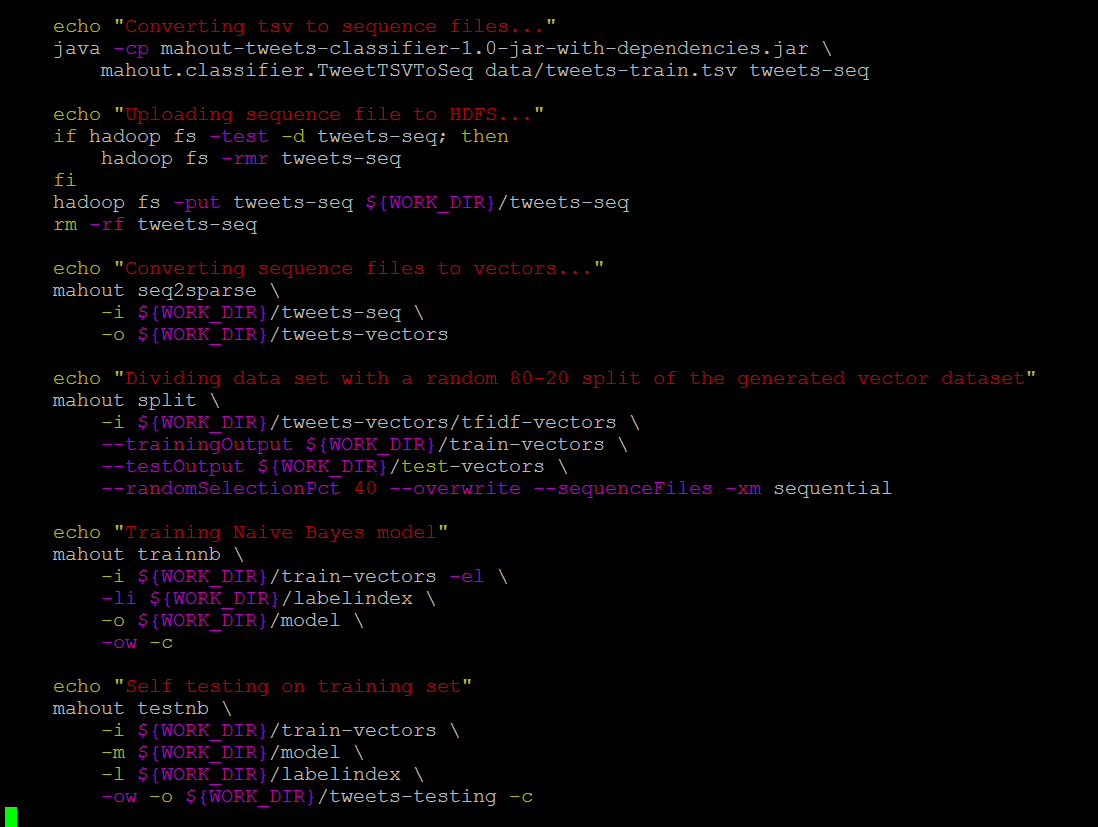
1. Performing sentimental analysis using python script:



1. Hive query extracting final report for tweets along with sentimental analysis:

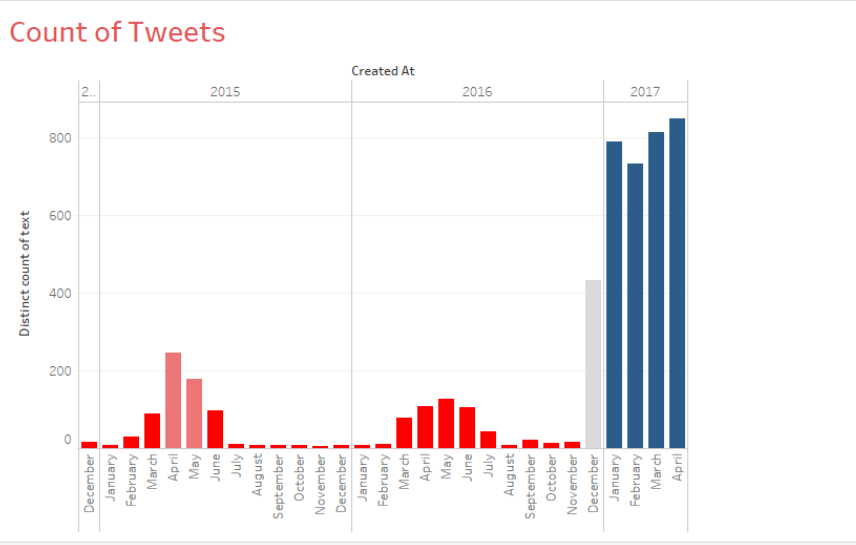


**Code Snippet:**

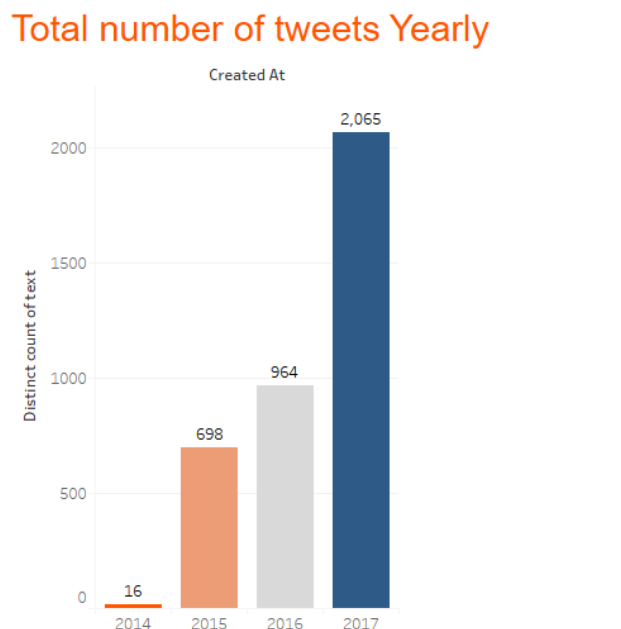


# Reporting and dashboard for data visualization:

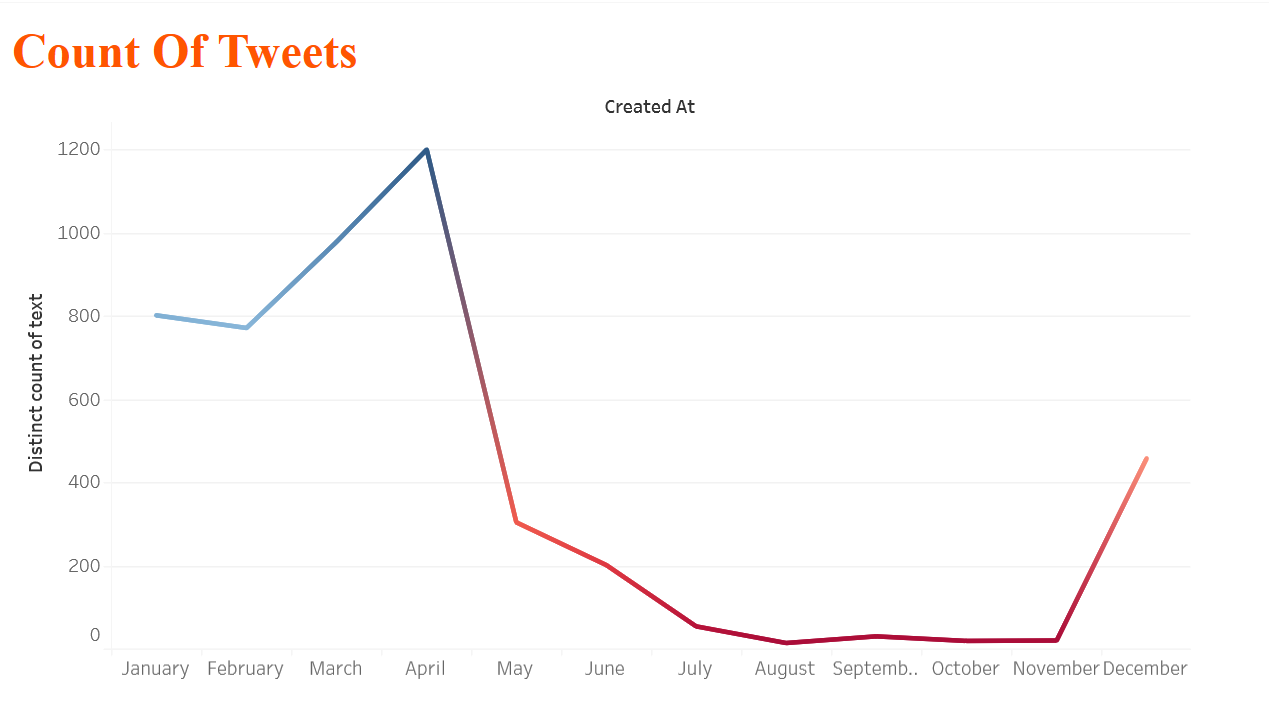
1. Number of tweets:



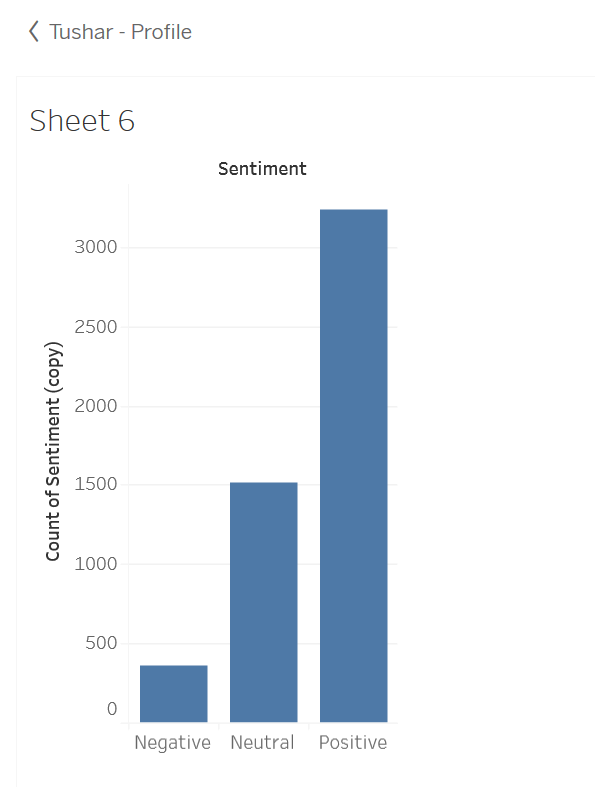
1. Total number of tweets Yearly:



1. Trendline



1. Sentiment count of tweets:



# References:

<https://chameerawijebandara.wordpress.com/2014/01/03/install-mahout-in-ubuntu-for-beginners/>

<http://hadooptutorial.info/apache-flume-installation/>

<https://www.edureka.co/blog/apache-hive-installation-on-ubuntu>

<http://apache.mirrors.pair.com/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz>

<http://www.nltk.org/howto/sentiment.html>