UMKC

Team 16 Project

PRINCIPLES OF BIG DATA MANAGEMENT

Namrata Dutta |nd8gv|16247052
Pujita Mullapudi|pm47c|16245822
Sravani Konujula|skv87|16230172
Sudheesha Reddy|smx75|16241536

Supervision:

Dr. Anas Katib

Introduction

Big data describes the large volume of both structured and unstructured data that is so large, it is difficult to process using traditional software techniques. Big Data has the potential to help companies improve operations and make faster, more intelligent decisions because of the competitive market. This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues and improve operations.

Hadoop is an open-source software framework used for distributed storage and processing of big data sets using the MapReduce programming model. Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster. HDFS is built to support applications with large data sets, including individual files that reach into the terabytes. It uses a master/slave architecture, with each cluster consisting of a single Name Node that manages file system operations and supporting Data Nodes that manage data storage on individual compute nodes.

Goal of the project

Collection of 100k tweets in JavaScript Object Notation (JSON) format
Find the list of top ten used hashtags in your collection.
Creation of a directory in HDFS for each hashtag from the top ten hashtag list.
Creation of 2 additional directories "Others" and "None"
Store the tweets on files in HDFS
- If a tweet contains a hashtag from the top ten list, store the tweet in that hashtag's directory.
-If a tweet contains one or more hashtags, but none of the hashtags are in the top ten list, store the tweet in the "Others" directory.
-If a tweet does not contain a hashtag, store it in the "None" directory.
Also, Implementation of a function that counts the number of times a keyword appears in one of two tweet JSON attributes (text and hashtags) in all of 12 directories that were created in HDFS

Tools Used

- **Python**
- **Ubuntu 14.04**
- **Hadoop 2.7.1**
- **Virtual box**
- **NetBeans**

Implementation

1. Collection of tweets using Python

   -Creation of Twitter Developer account to get access tokens

   -Using Python, we collected the tweets in JSON format with keywords like “Cricket”, “Obama”, “Oscars2017”

   -Converted the tweets into text files
2. Installation of Virtual box and Ubuntu to set the environment
3. Installation of Hadoop

   - Installed Java

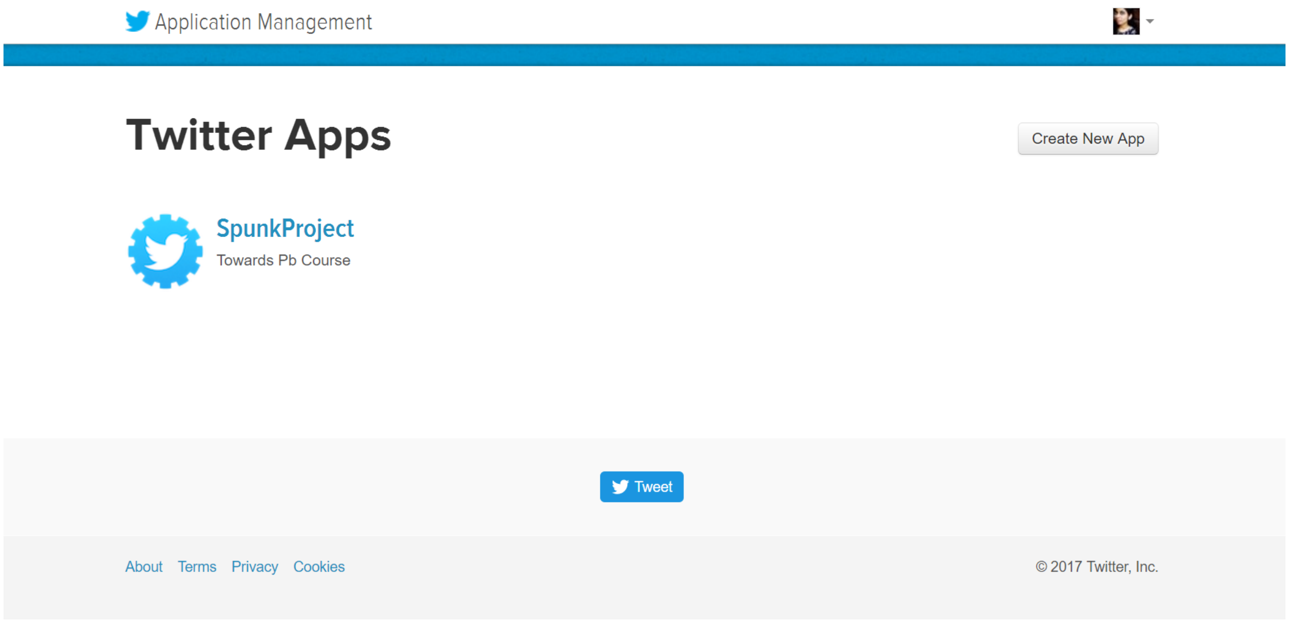
   -Installed Scala

   -Installed Spark
4. Start Hadoop
5. Creating directories and store in HDFS files

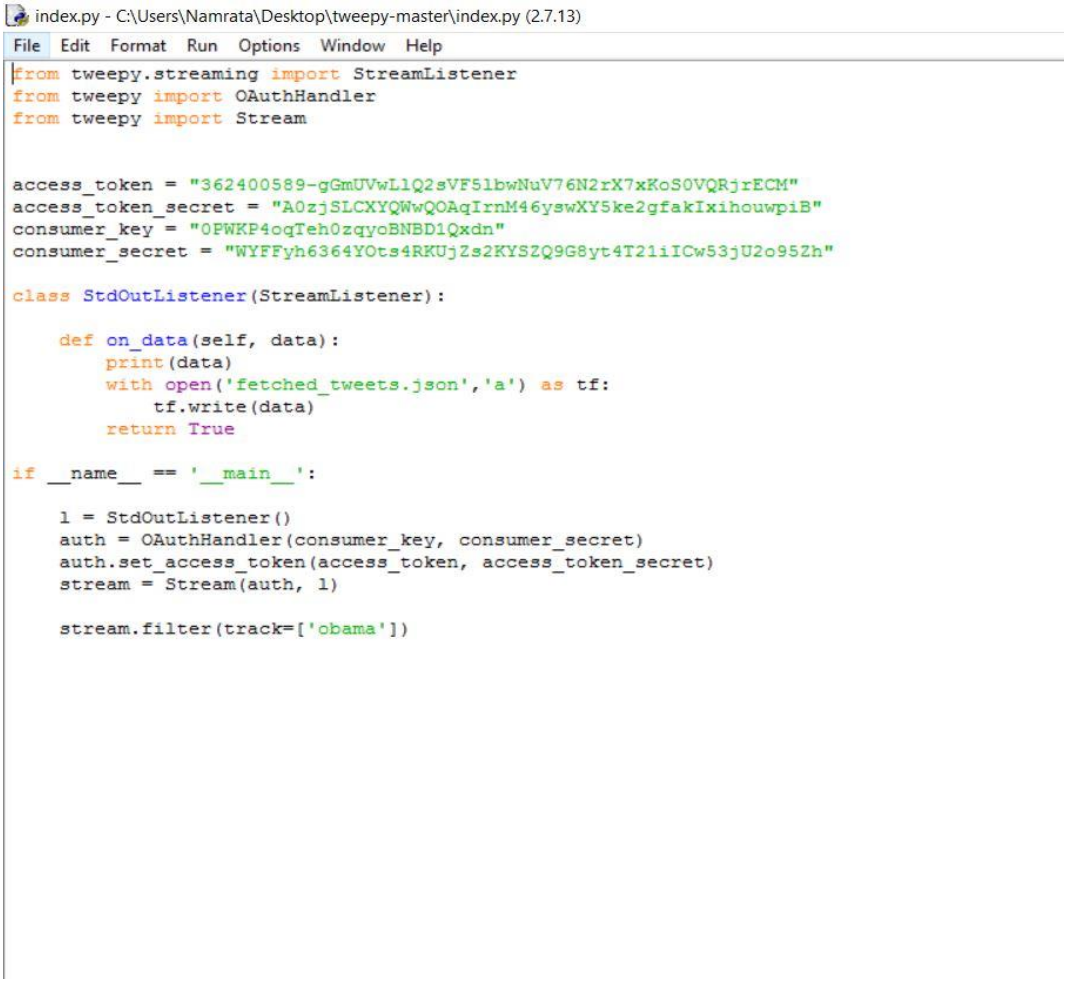
Output Screenshots

1. Collection of Tweets using Python

-Creation of Dev account to get access tokens



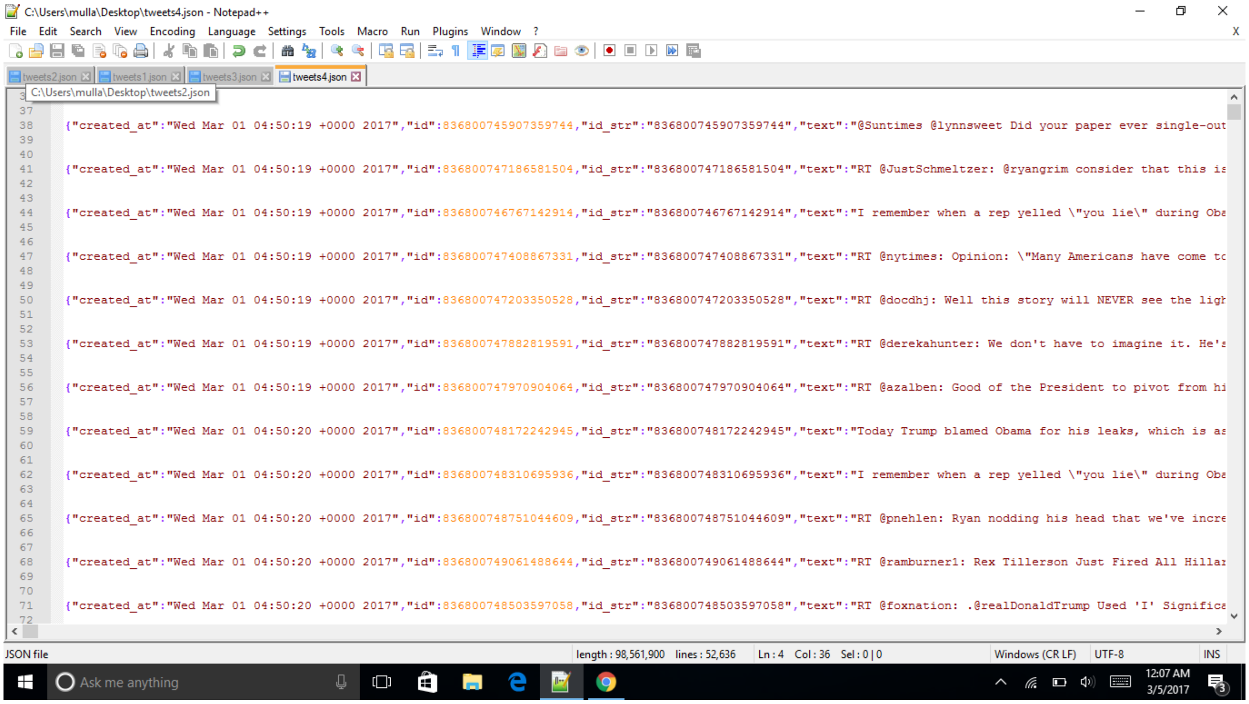
-Python program for tweets collection



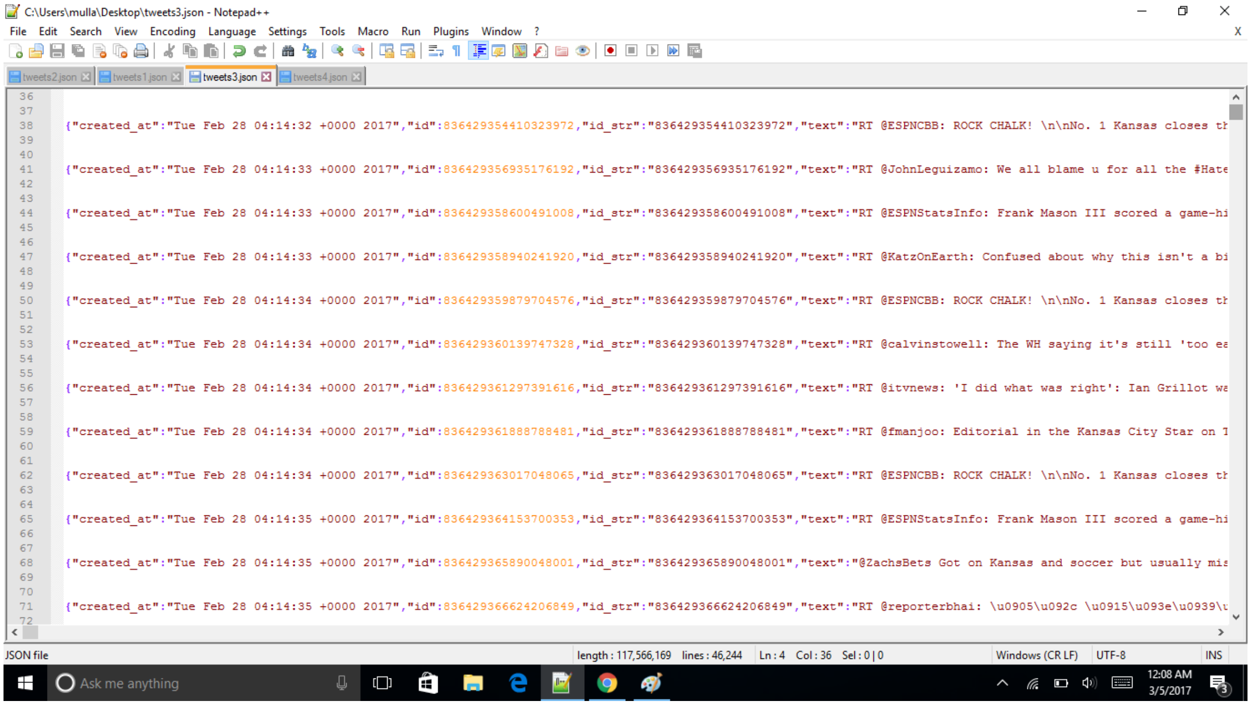
index.py - C:\Users\Namrata\Desktop\tweepy-master\index.py (2.7.13)
File Edit Format Run Options Window Help

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream


access_token = "362400589-gGmUVwLlQ2sVF5lbwNuV76N2rX7xKoS0VQRjrECM"
access_token_secret = "A0zjSLCXYQWwQOAqIrnM46yswXY5ke2gfakIxihouwpiB"
consumer_key = "0PWKP4oqTeh0zqyoBNBD1Qxdn"
consumer_secret = "WYFFyh6364YOts4RKUjZs2KYSZQ9G8yt4T21iICw53jU2o95Zh"

class StdOutListener(StreamListener):

    def on_data(self, data):
        print(data)
        with open('fetched_tweets.json','a') as tf:
            tf.write(data)
        return True

if __name__ == '__main__':

    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    stream.filter(track=['obama'])
```

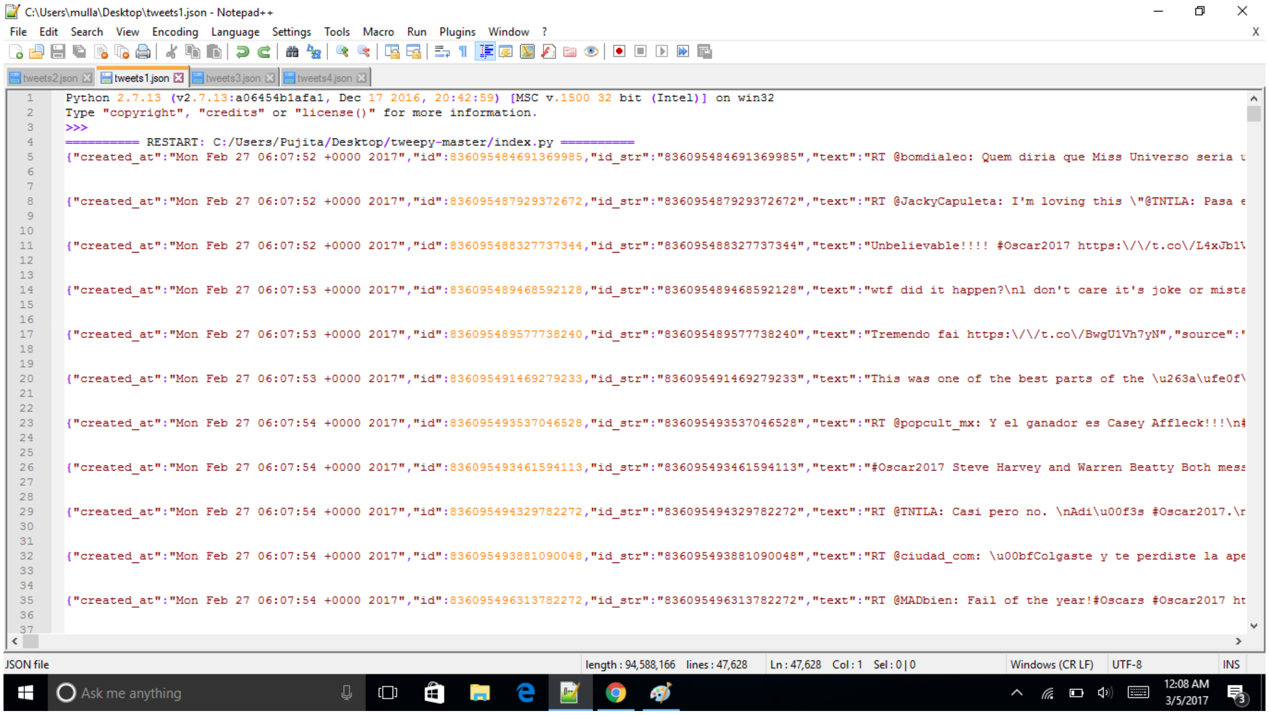
- Tweets Collected


C:\Users\mulla\Desktop\tweets4.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
tweets2.json
tweets1.json
tweets3.json
tweets4.json
C:\Users\mulla\Desktop\tweets2.json
JSON file
length : 98,561,900 lines : 52,636
Ln : 4 Col : 36 Sel : 0 | 0
Windows (CR LF) UTF-8 INS
Ask me anything
12:07 AM
3/5/2017

```
{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800745907359744,"id_str":"836800745907359744","text":"@Suntimes @lynnsweet Did your paper ever single-out

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800747186581504,"id_str":"836800747186581504","text":"RT @JustSchmeltzer: @ryangrim consider that this is

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800746767142914,"id_str":"836800746767142914","text":"I remember when a rep yelled \"you lie\" during Oba

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800747408867331,"id_str":"836800747408867331","text":"RT @nytimes: Opinion: \"Many Americans have come tc

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800747203350528,"id_str":"836800747203350528","text":"RT @docdhj: Well this story will NEVER see the ligh

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800747882819591,"id_str":"836800747882819591","text":"RT @derekahunter: We don't have to imagine it. He's

{"created_at":"Wed Mar 01 04:50:19 +0000 2017","id":836800747970904064,"id_str":"836800747970904064","text":"RT @azalben: Good of the President to pivot from hi

{"created_at":"Wed Mar 01 04:50:20 +0000 2017","id":836800748172242945,"id_str":"836800748172242945","text":"Today Trump blamed Obama for his leaks, which is as

{"created_at":"Wed Mar 01 04:50:20 +0000 2017","id":836800748310695936,"id_str":"836800748310695936","text":"I remember when a rep yelled \"you lie\" during Oba

{"created_at":"Wed Mar 01 04:50:20 +0000 2017","id":836800748751044609,"id_str":"836800748751044609","text":"RT @pnehlen: Ryan nodding his head that we've incre

{"created_at":"Wed Mar 01 04:50:20 +0000 2017","id":836800749061488644,"id_str":"836800749061488644","text":"RT @ramburner1: Rex Tillerson Just Fired All Hillar

{"created_at":"Wed Mar 01 04:50:20 +0000 2017","id":836800748503597058,"id_str":"836800748503597058","text":"RT @foxnation: .@realDonaldTrump Used 'I' Significa
```


C:\Users\mulla\Desktop\tweets3.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
tweets2.json
tweets1.json
tweets3.json
tweets4.json
JSON file
length : 117,566,169 lines : 46,244
Ln : 4 Col : 36 Sel : 0 | 0
Windows (CR LF) UTF-8 INS
Ask me anything
12:08 AM
3/5/2017

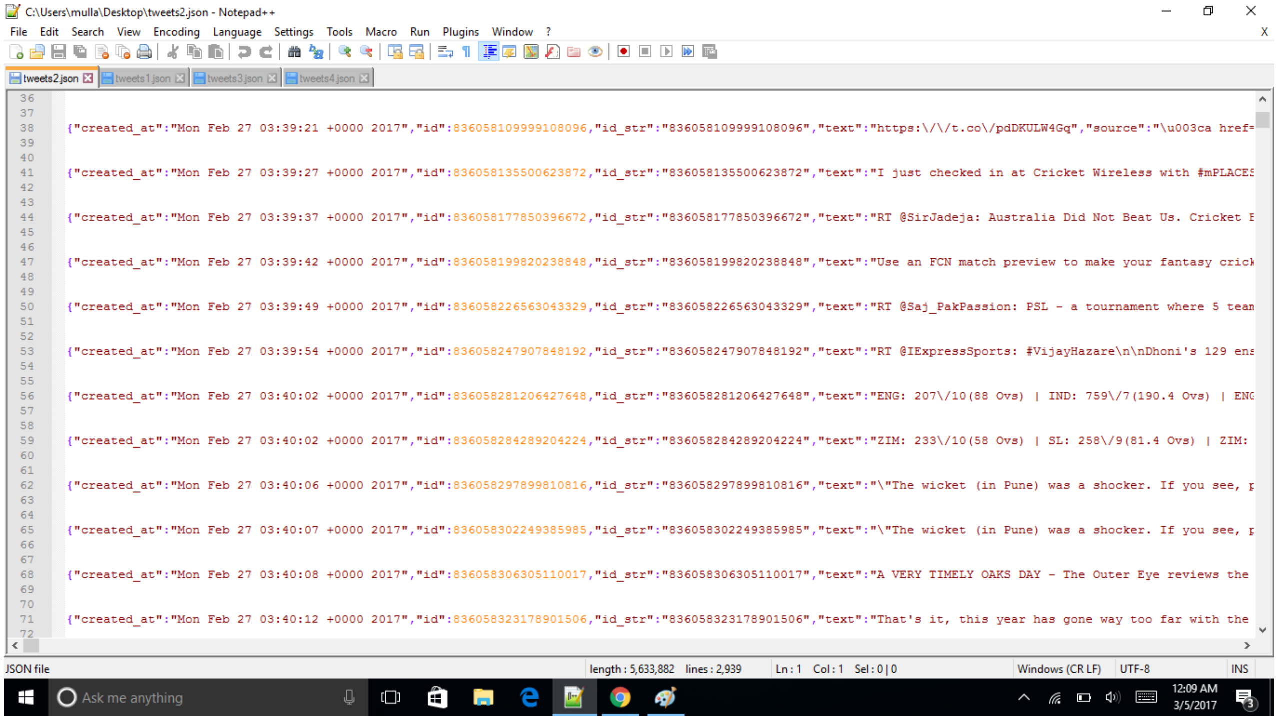
```
{"created_at":"Tue Feb 28 04:14:32 +0000 2017","id":836429354410323972,"id_str":"836429354410323972","text":"RT @ESPNCBB: ROCK CHALK! \n\nNo. 1 Kansas closes th

{"created_at":"Tue Feb 28 04:14:33 +0000 2017","id":836429356935176192,"id_str":"836429356935176192","text":"RT @JohnLeguizamo: We all blame u for all the #Hate

{"created_at":"Tue Feb 28 04:14:33 +0000 2017","id":836429358600491008,"id_str":"836429358600491008","text":"RT @ESPNStatsInfo: Frank Mason III scored a game-hi

{"created_at":"Tue Feb 28 04:14:33 +0000 2017","id":836429358940241920,"id_str":"836429358940241920","text":"RT @KatzOnEarth: Confused about why this isn't a bi

{"created_at":"Tue Feb 28 04:14:34 +0000 2017","id":836429359879704576,"id_str":"836429359879704576","text":"RT @ESPNCBB: ROCK CHALK! \n\nNo. 1 Kansas closes th

{"created_at":"Tue Feb 28 04:14:34 +0000 2017","id":836429360139747328,"id_str":"836429360139747328","text":"RT @calvinstowell: The WH saying it's still 'too ea

{"created_at":"Tue Feb 28 04:14:34 +0000 2017","id":836429361297391616,"id_str":"836429361297391616","text":"RT @itvnews: 'I did what was right': Ian Grillot wa

{"created_at":"Tue Feb 28 04:14:34 +0000 2017","id":836429361888788481,"id_str":"836429361888788481","text":"RT @fmanjoo: Editorial in the Kansas City Star on T

{"created_at":"Tue Feb 28 04:14:34 +0000 2017","id":836429363017048065,"id_str":"836429363017048065","text":"RT @ESPNCBB: ROCK CHALK! \n\nNo. 1 Kansas closes th

{"created_at":"Tue Feb 28 04:14:35 +0000 2017","id":836429364153700353,"id_str":"836429364153700353","text":"RT @ESPNStatsInfo: Frank Mason III scored a game-hi

{"created_at":"Tue Feb 28 04:14:35 +0000 2017","id":836429365890048001,"id_str":"836429365890048001","text":"@ZachsBets Got on Kansas and soccer but usually mis

{"created_at":"Tue Feb 28 04:14:35 +0000 2017","id":836429366624206849,"id_str":"836429366624206849","text":"RT @reporterbhai: \u0905\u092c \u0915\u093e\u0939\u
```


C:\Users\mulla\Desktop\tweets1.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
tweets2.json
tweets1.json
tweets3.json
tweets4.json
Ask me anything
12:08 AM
3/5/2017

```
Python 2.7.13 (v2.7.13:a06454b1afa1, Dec 17 2016, 20:42:59) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:/Users/Pujita/Desktop/tweepy-master/index.py ==========
```

```
{"created_at":"Mon Feb 27 06:07:52 +0000 2017","id":836095484691369985,"id_str":"836095484691369985","text":"RT @bomdialeo: Quem diria que Miss Universo seria u


{"created_at":"Mon Feb 27 06:07:52 +0000 2017","id":836095487929372672,"id_str":"836095487929372672","text":"RT @JackyCapuleta: I'm loving this \"@TNTLA: Pasa e


{"created_at":"Mon Feb 27 06:07:52 +0000 2017","id":836095488327737344,"id_str":"836095488327737344","text":"Unbelievable!!!! #Oscar2017 https:\/\/t.co\/L4xJb1V


{"created_at":"Mon Feb 27 06:07:53 +0000 2017","id":836095489468592128,"id_str":"836095489468592128","text":"wtf did it happen?\nl don't care it's joke or mista


{"created_at":"Mon Feb 27 06:07:53 +0000 2017","id":836095489577738240,"id_str":"836095489577738240","text":"Tremendo fai https:\/\/t.co\/BwgU1Vh7yN","source":"


{"created_at":"Mon Feb 27 06:07:53 +0000 2017","id":836095491469279233,"id_str":"836095491469279233","text":"This was one of the best parts of the \u263a\ufe0f\


{"created_at":"Mon Feb 27 06:07:54 +0000 2017","id":836095493537046528,"id_str":"836095493537046528","text":"RT @popcult_mx: Y el ganador es Casey Affleck!!!\n#


{"created_at":"Mon Feb 27 06:07:54 +0000 2017","id":836095493461594113,"id_str":"836095493461594113","text":"#Oscar2017 Steve Harvey and Warren Beatty Both mess


{"created_at":"Mon Feb 27 06:07:54 +0000 2017","id":836095494329782272,"id_str":"836095494329782272","text":"RT @TNTLA: Casi pero no. \nAdi\u00f3s #Oscar2017.\r


{"created_at":"Mon Feb 27 06:07:54 +0000 2017","id":836095493881090048,"id_str":"836095493881090048","text":"RT @ciudad_com: \u00bfColgaste y te perdiste la ape


{"created_at":"Mon Feb 27 06:07:54 +0000 2017","id":836095496313782272,"id_str":"836095496313782272","text":"RT @MADbien: Fail of the year!#Oscars #Oscar2017 ht
```

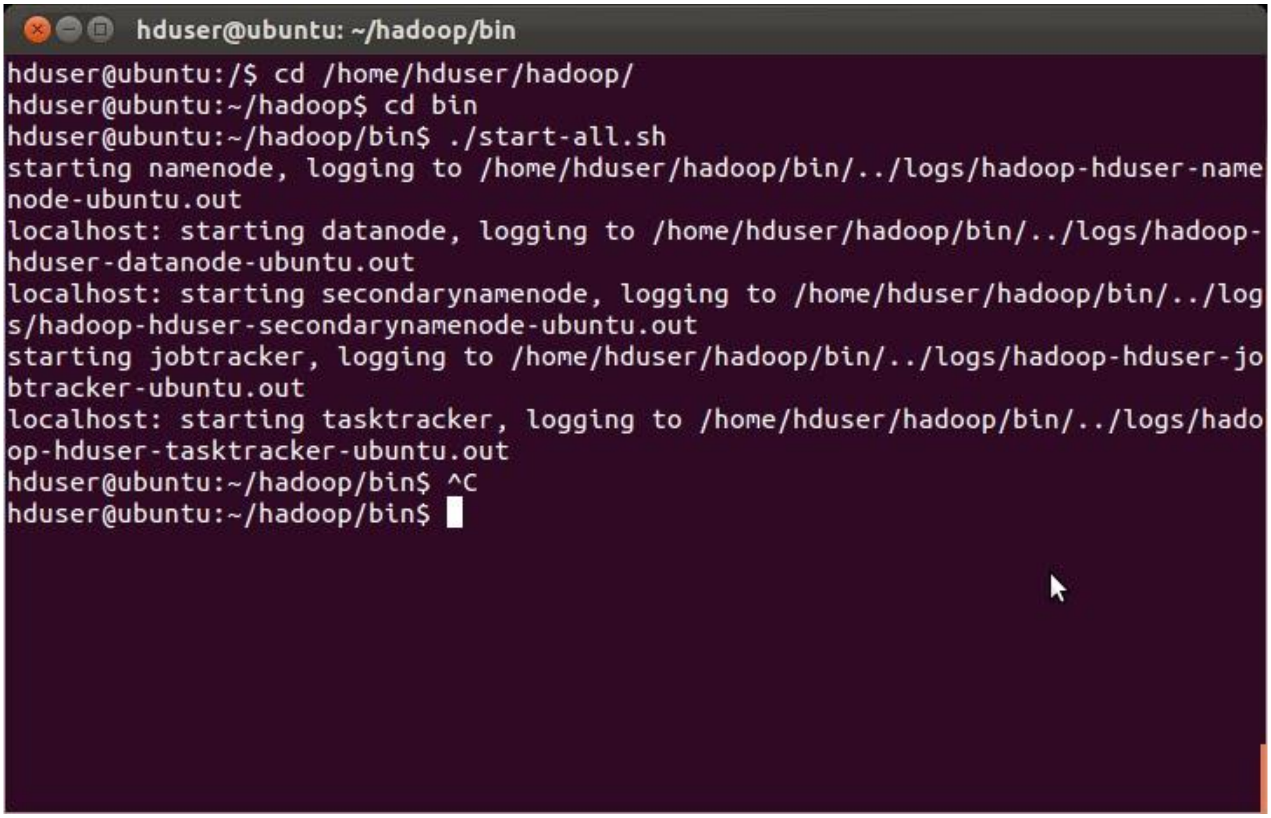
JSON file | length : 94,588,166 lines : 47,628 | Ln : 47,628 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS


C:\Users\mulla\Desktop\tweets2.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
tweets2.json
tweets1.json
tweets3.json
tweets4.json
Ask me anything
12:09 AM
3/5/2017

```


{"created_at":"Mon Feb 27 03:39:21 +0000 2017","id":836058109999108096,"id_str":"836058109999108096","text":"https:\/\/t.co\/pdDKULW4Gq","source":"\u003ca href=


{"created_at":"Mon Feb 27 03:39:27 +0000 2017","id":836058135500623872,"id_str":"836058135500623872","text":"I just checked in at Cricket Wireless with #mPLACES


{"created_at":"Mon Feb 27 03:39:37 +0000 2017","id":836058177850396672,"id_str":"836058177850396672","text":"RT @SirJadeja: Australia Did Not Beat Us. Cricket E


{"created_at":"Mon Feb 27 03:39:42 +0000 2017","id":836058199820238848,"id_str":"836058199820238848","text":"Use an FCN match preview to make your fantasy crick


{"created_at":"Mon Feb 27 03:39:49 +0000 2017","id":836058226563043329,"id_str":"836058226563043329","text":"RT @Saj_PakPassion: PSL - a tournament where 5 team


{"created_at":"Mon Feb 27 03:39:54 +0000 2017","id":836058247907848192,"id_str":"836058247907848192","text":"RT @IExpressSports: #VijayHazare\n\nDhoni's 129 ens


{"created_at":"Mon Feb 27 03:40:02 +0000 2017","id":836058281206427648,"id_str":"836058281206427648","text":"ENG: 207\/10(88 Ovs) | IND: 759\/7(190.4 Ovs) | ENG


{"created_at":"Mon Feb 27 03:40:02 +0000 2017","id":836058284289204224,"id_str":"836058284289204224","text":"ZIM: 233\/10(58 Ovs) | SL: 258\/9(81.4 Ovs) | ZIM:


{"created_at":"Mon Feb 27 03:40:06 +0000 2017","id":836058297899810816,"id_str":"836058297899810816","text":"\"The wicket (in Pune) was a shocker. If you see, p


{"created_at":"Mon Feb 27 03:40:07 +0000 2017","id":836058302249385985,"id_str":"836058302249385985","text":"\"The wicket (in Pune) was a shocker. If you see, p


{"created_at":"Mon Feb 27 03:40:08 +0000 2017","id":836058306305110017,"id_str":"836058306305110017","text":"A VERY TIMELY OAKS DAY - The Outer Eye reviews the


{"created_at":"Mon Feb 27 03:40:12 +0000 2017","id":836058323178901506,"id_str":"836058323178901506","text":"That's it, this year has gone way too far with the
```

JSON file | length : 5,633,882 lines : 2,939 | Ln : 1 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

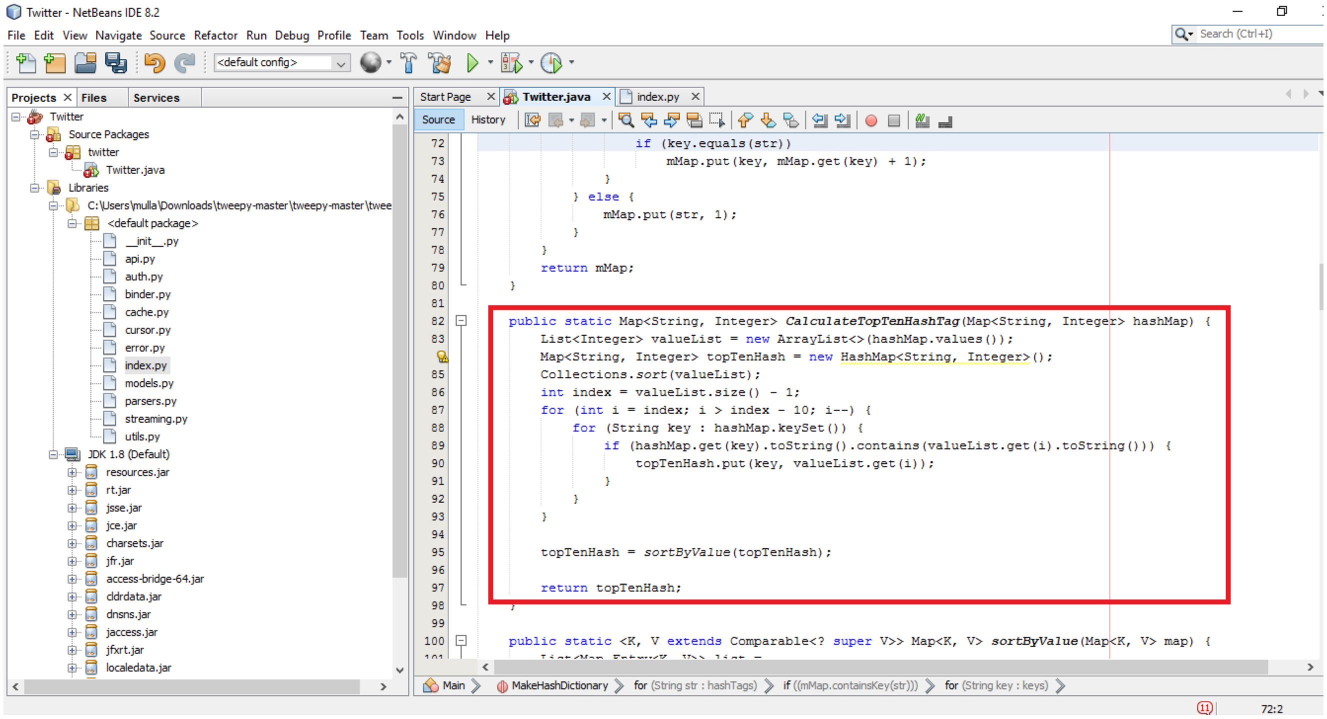
2. Start Hadoop



```
hduser@ubuntu:/$ cd /home/hduser/hadoop/
hduser@ubuntu:~/hadoop$ cd bin
hduser@ubuntu:~/hadoop/bin$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/bin/../logs/hadoop-hduser-name
node-ubuntu.out
localhost: starting datanode, logging to /home/hduser/hadoop/bin/../logs/hadoop-
hduser-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/bin/../log
s/hadoop-hduser-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/hduser/hadoop/bin/../logs/hadoop-hduser-jo
btracker-ubuntu.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/bin/../logs/hado
op-hduser-tasktracker-ubuntu.out
hduser@ubuntu:~/hadoop/bin$ ^C
hduser@ubuntu:~/hadoop/bin$ 
```

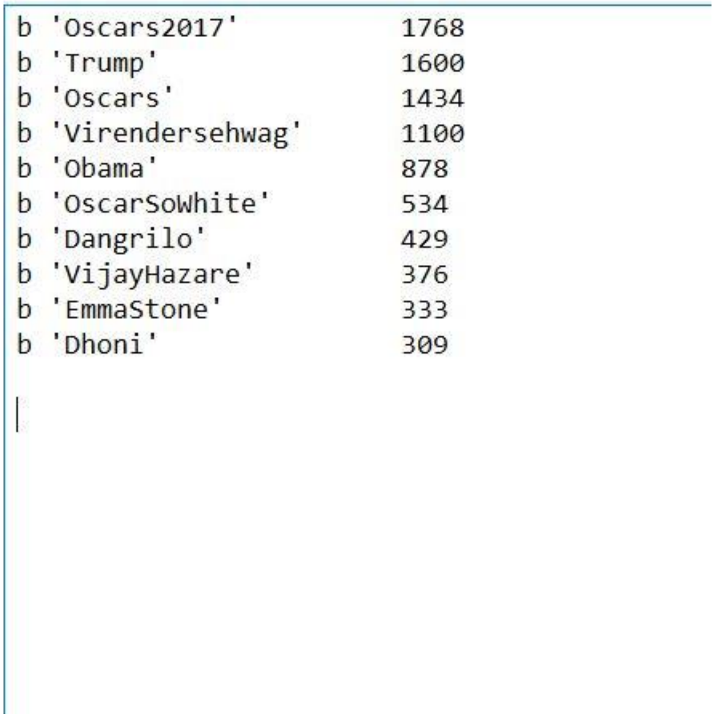
3. Top Ten Hashtags

- Program for top 10 hashtags



```java
                if (key.equals(str))
                    mMap.put(key, mMap.get(key) + 1);
            }
        } else {
            mMap.put(str, 1);
        }
    }
    return mMap;
}

public static Map<String, Integer> CalculateTopTenHashTag(Map<String, Integer> hashMap) {
    List<Integer> valueList = new ArrayList<>(hashMap.values());
    Map<String, Integer> topTenHash = new HashMap<String, Integer>();
    Collections.sort(valueList);
    int index = valueList.size() - 1;
    for (int i = index; i > index - 10; i--) {
        for (String key : hashMap.keySet()) {
            if (hashMap.get(key).toString().contains(valueList.get(i).toString())) {
                topTenHash.put(key, valueList.get(i));
            }
        }
    }

    topTenHash = sortByValue(topTenHash);

    return topTenHash;
}

public static <K, V extends Comparable<? super V>> Map<K, V> sortByValue(Map<K, V> map) {
```

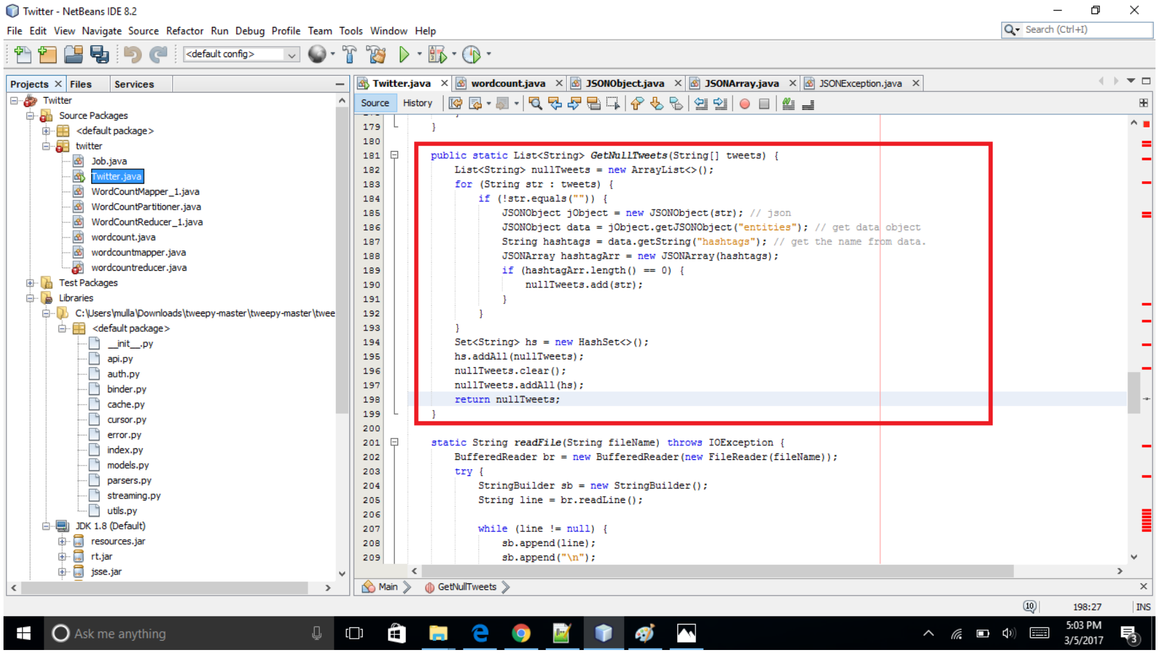
- Output of Top 10 Hashtags



```
b 'Oscars2017'        1768
b 'Trump'             1600
b 'Oscars'            1434
b 'Virendersehwag'    1100
b 'Obama'             878
b 'OscarSoWhite'      534
b 'Dangrilo'          429
b 'VijayHazare'       376
b 'EmmaStone'         333
b 'Dhoni'             309
```

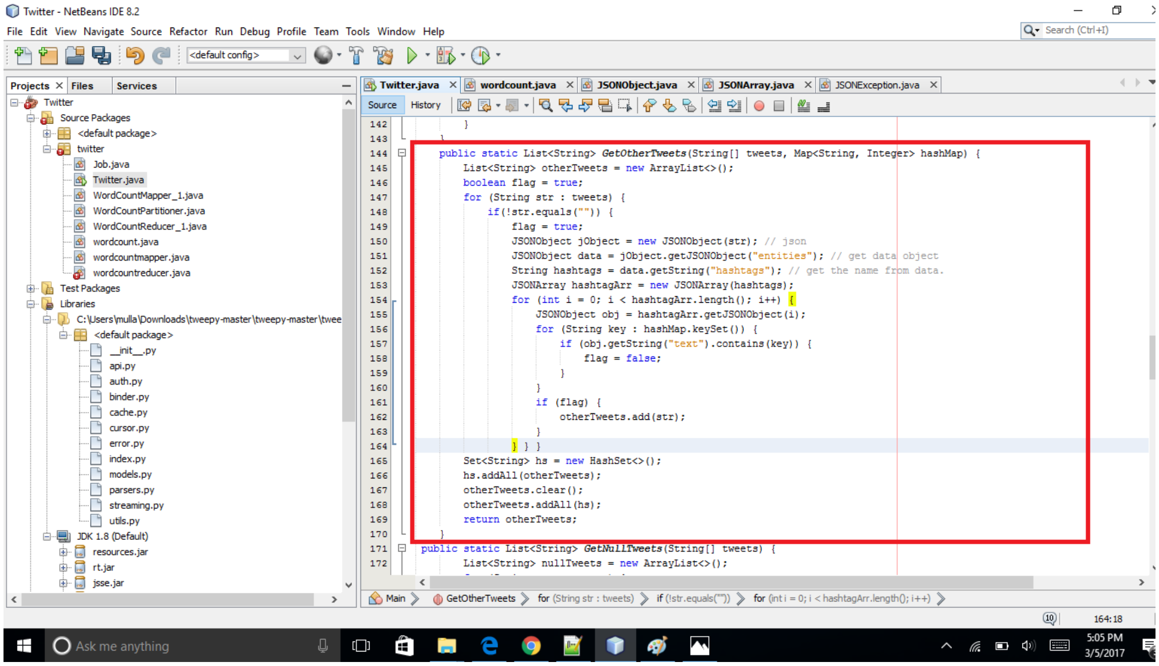
4. Program for creating directories

- Null tweets


Twitter - NetBeans IDE 8.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Projects Files Services
Twitter.java wordcount.java JSONObject.java JSONArray.java JSONException.java
Source History
Main GetNullTweets
198:27 INS
Ask me anything
5:03 PM 3/5/2017

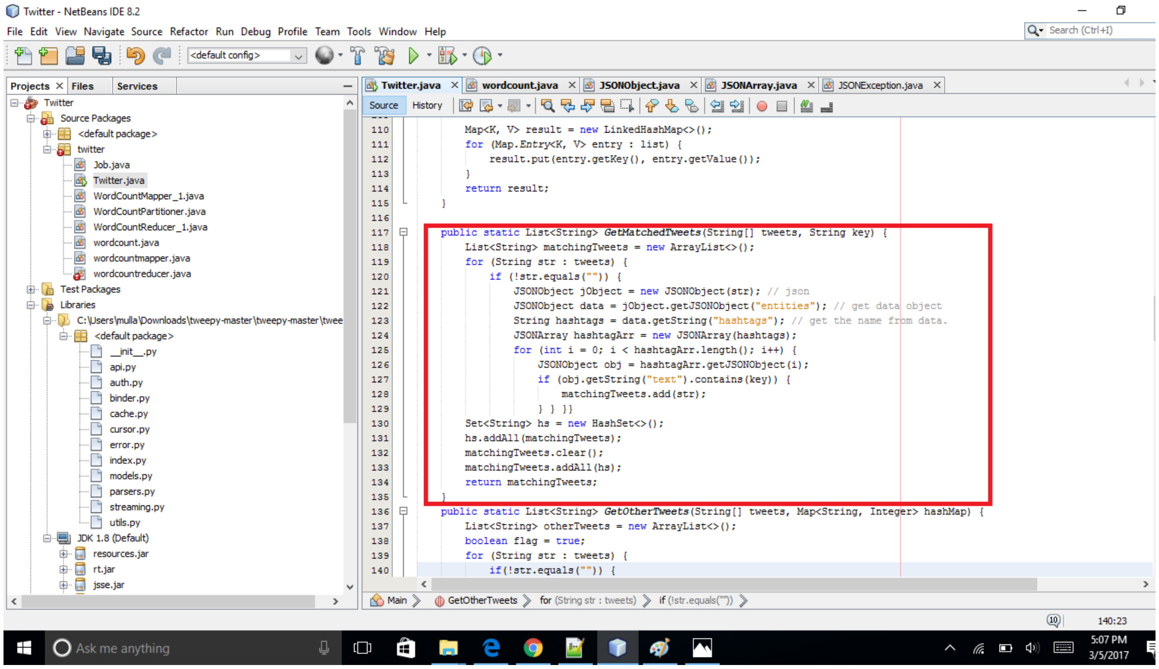
```
    }

    public static List<String> GetNullTweets(String[] tweets) {
        List<String> nullTweets = new ArrayList<>();
        for (String str : tweets) {
            if (!str.equals("")) {
                JSONObject jObject = new JSONObject(str); // json
                JSONObject data = jObject.getJSONObject("entities"); // get data object
                String hashtags = data.getString("hashtags"); // get the name from data.
                JSONArray hashtagArr = new JSONArray(hashtags);
                if (hashtagArr.length() == 0) {
                    nullTweets.add(str);
                }
            }
        }
        Set<String> hs = new HashSet<>();
        hs.addAll(nullTweets);
        nullTweets.clear();
        nullTweets.addAll(hs);
        return nullTweets;
    }

    static String readFile(String fileName) throws IOException {
        BufferedReader br = new BufferedReader(new FileReader(fileName));
        try {
            StringBuilder sb = new StringBuilder();
            String line = br.readLine();

            while (line != null) {
                sb.append(line);
                sb.append("\n");
```

- Other Tweets


Twitter - NetBeans IDE 8.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Projects Files Services
Twitter.java wordcount.java JSONObject.java JSONArray.java JSONException.java
Source History
Main GetOtherTweets for (String str : tweets) if (!str.equals("")) for (int i = 0; i < hashtagArr.length(); i++)
164:18
Ask me anything
5:05 PM 3/5/2017

```
            }
    }
    public static List<String> GetOtherTweets(String[] tweets, Map<String, Integer> hashMap) {
        List<String> otherTweets = new ArrayList<>();
        boolean flag = true;
        for (String str : tweets) {
            if(!str.equals("")) {
                flag = true;
                JSONObject jObject = new JSONObject(str); // json
                JSONObject data = jObject.getJSONObject("entities"); // get data object
                String hashtags = data.getString("hashtags"); // get the name from data.
                JSONArray hashtagArr = new JSONArray(hashtags);
                for (int i = 0; i < hashtagArr.length(); i++) {
                    JSONObject obj = hashtagArr.getJSONObject(i);
                    for (String key : hashMap.keySet()) {
                        if (obj.getString("text").contains(key)) {
                            flag = false;
                        }
                    }
                    if (flag) {
                        otherTweets.add(str);
                    }
                } } }
        Set<String> hs = new HashSet<>();
        hs.addAll(otherTweets);
        otherTweets.clear();
        otherTweets.addAll(hs);
        return otherTweets;
    }
 public static List<String> GetNullTweets(String[] tweets) {
        List<String> nullTweets = new ArrayList<>();
```

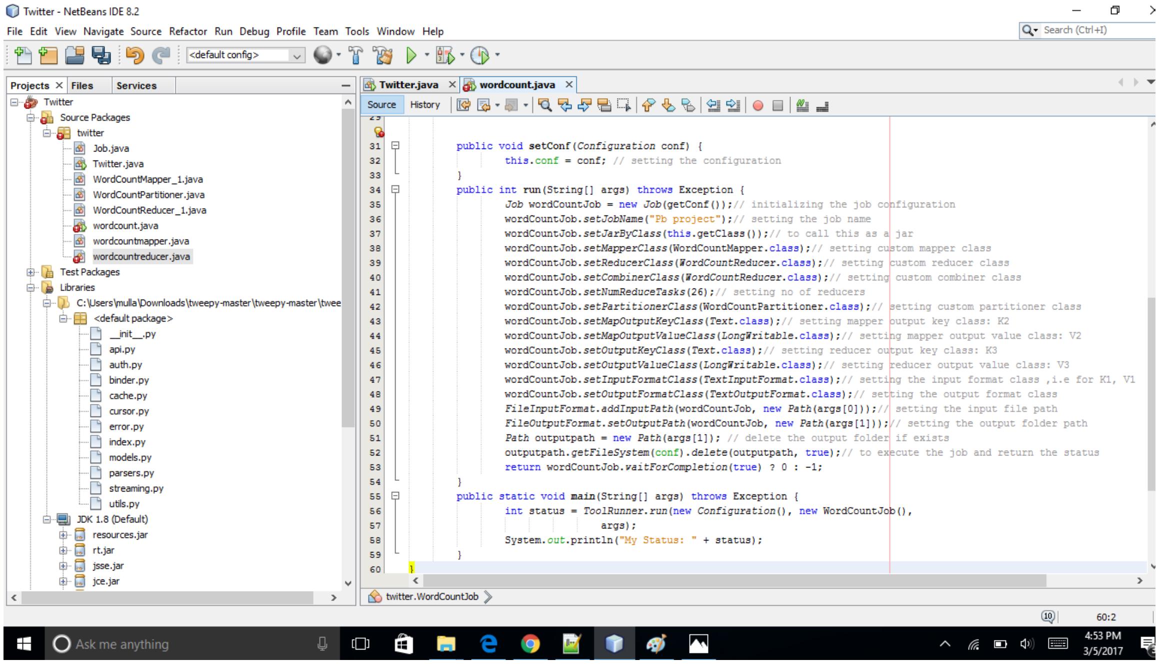
- Matched Tweets



```java
        Map<K, V> result = new LinkedHashMap<>();
        for (Map.Entry<K, V> entry : list) {
            result.put(entry.getKey(), entry.getValue());
        }
        return result;
    }

    public static List<String> GetMatchedTweets(String[] tweets, String key) {
        List<String> matchingTweets = new ArrayList<>();
        for (String str : tweets) {
            if (!str.equals("")) {
                JSONObject jObject = new JSONObject(str); // json
                JSONObject data = jObject.getJSONObject("entities"); // get data object
                String hashtags = data.getString("hashtags"); // get the name from data.
                JSONArray hashtagArr = new JSONArray(hashtags);
                for (int i = 0; i < hashtagArr.length(); i++) {
                    JSONObject obj = hashtagArr.getJSONObject(i);
                    if (obj.getString("text").contains(key)) {
                        matchingTweets.add(str);
                    } } }}
        Set<String> hs = new HashSet<>();
        hs.addAll(matchingTweets);
        matchingTweets.clear();
        matchingTweets.addAll(hs);
        return matchingTweets;
    }
    public static List<String> GetOtherTweets(String[] tweets, Map<String, Integer> hashMap) {
        List<String> otherTweets = new ArrayList<>();
        boolean flag = true;
        for (String str : tweets) {
            if(!str.equals("")) {
```

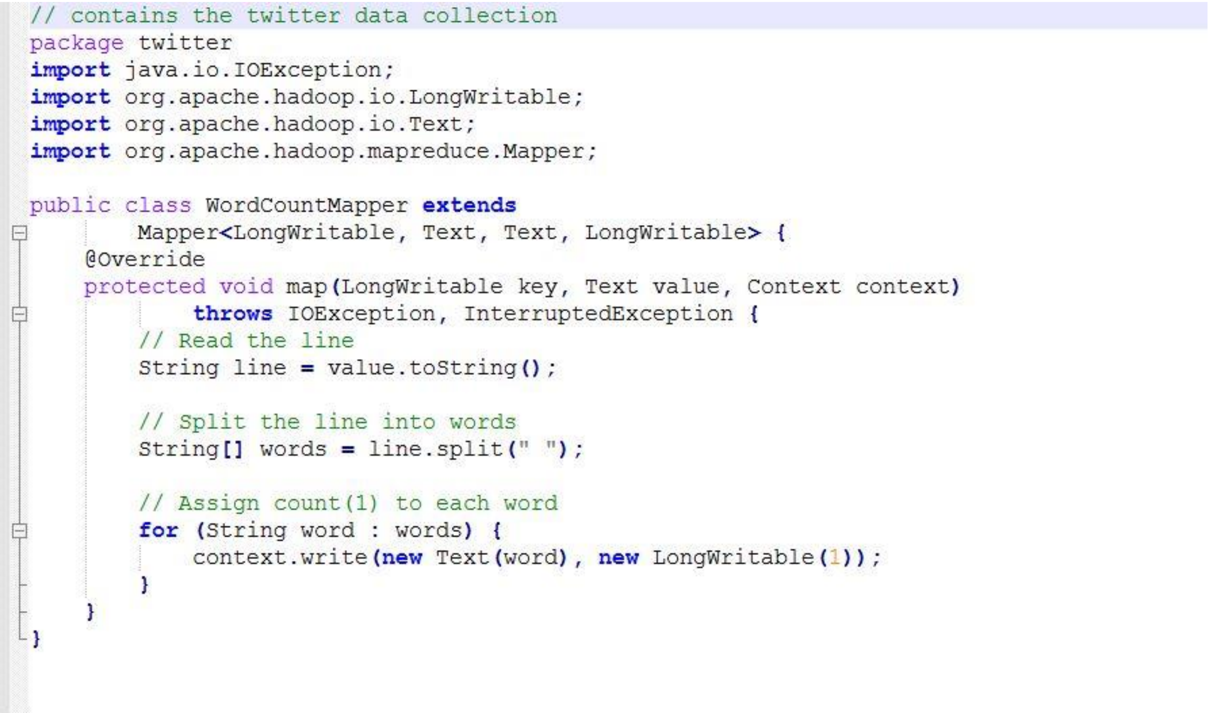
5. Extra Requirement: Word Count

- Word Count Job Program



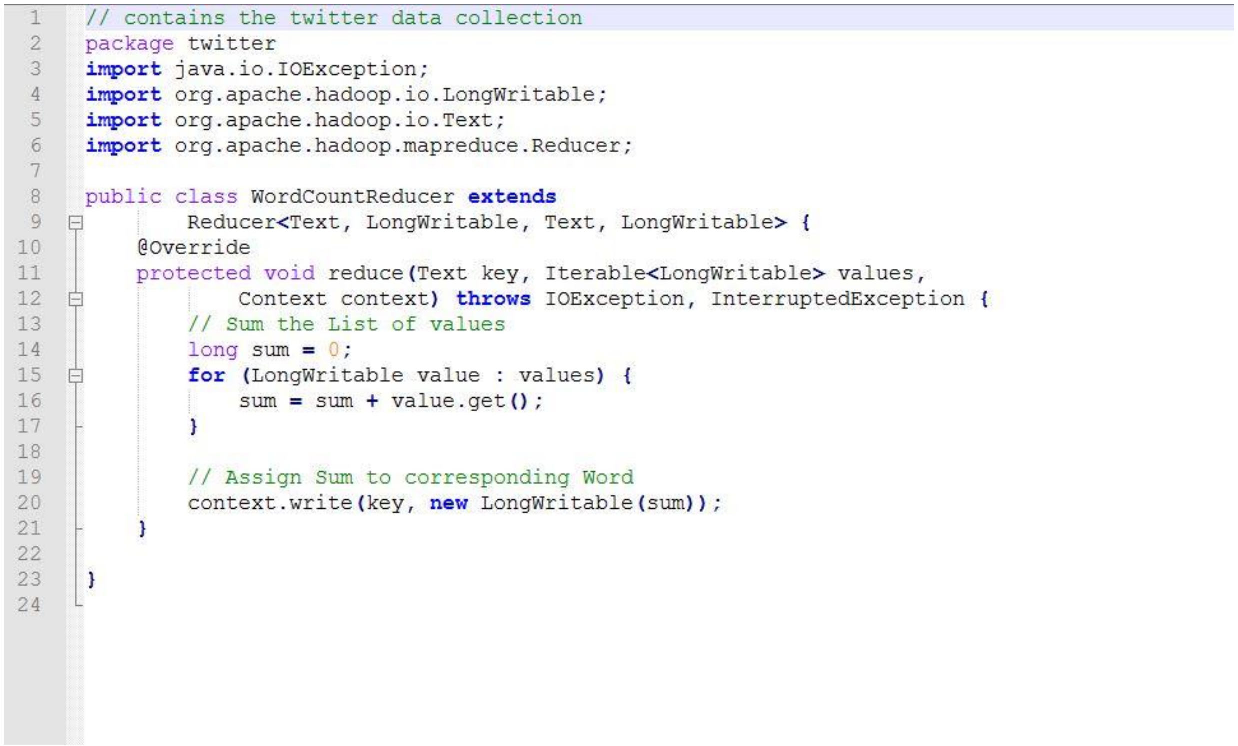
```java
    public void setConf(Configuration conf) {
            this.conf = conf; // setting the configuration
    }
    public int run(String[] args) throws Exception {
            Job wordCountJob = new Job(getConf());// initializing the job configuration
            wordCountJob.setJobName("Pb project");// setting the job name
            wordCountJob.setJarByClass(this.getClass());// to call this as a jar
            wordCountJob.setMapperClass(WordCountMapper.class);// setting custom mapper class
            wordCountJob.setReducerClass(WordCountReducer.class);// setting custom reducer class
            wordCountJob.setCombinerClass(WordCountReducer.class);// setting custom combiner class
            wordCountJob.setNumReduceTasks(26);// setting no of reducers
            wordCountJob.setPartitionerClass(WordCountPartitioner.class);// setting custom partitioner class
            wordCountJob.setMapOutputKeyClass(Text.class);// setting mapper output key class: K2
            wordCountJob.setMapOutputValueClass(LongWritable.class);// setting mapper output value class: V2
            wordCountJob.setOutputKeyClass(Text.class);// setting reducer output key class: K3
            wordCountJob.setOutputValueClass(LongWritable.class);// setting reducer output value class: V3
            wordCountJob.setInputFormatClass(TextInputFormat.class);// setting the input format class ,i.e for K1, V1
            wordCountJob.setOutputFormatClass(TextOutputFormat.class);// setting the output format class
            FileInputFormat.addInputPath(wordCountJob, new Path(args[0]));// setting the input file path
            FileOutputFormat.setOutputPath(wordCountJob, new Path(args[1]));// setting the output folder path
            Path outputpath = new Path(args[1]); // delete the output folder if exists
            outputpath.getFileSystem(conf).delete(outputpath, true);// to execute the job and return the status
            return wordCountJob.waitForCompletion(true) ? 0 : -1;
    }
    public static void main(String[] args) throws Exception {
            int status = ToolRunner.run(new Configuration(), new WordCountJob(),
                                args);
            System.out.println("My Status: " + status);
    }
}
```

- Word Count Mapper Function Program



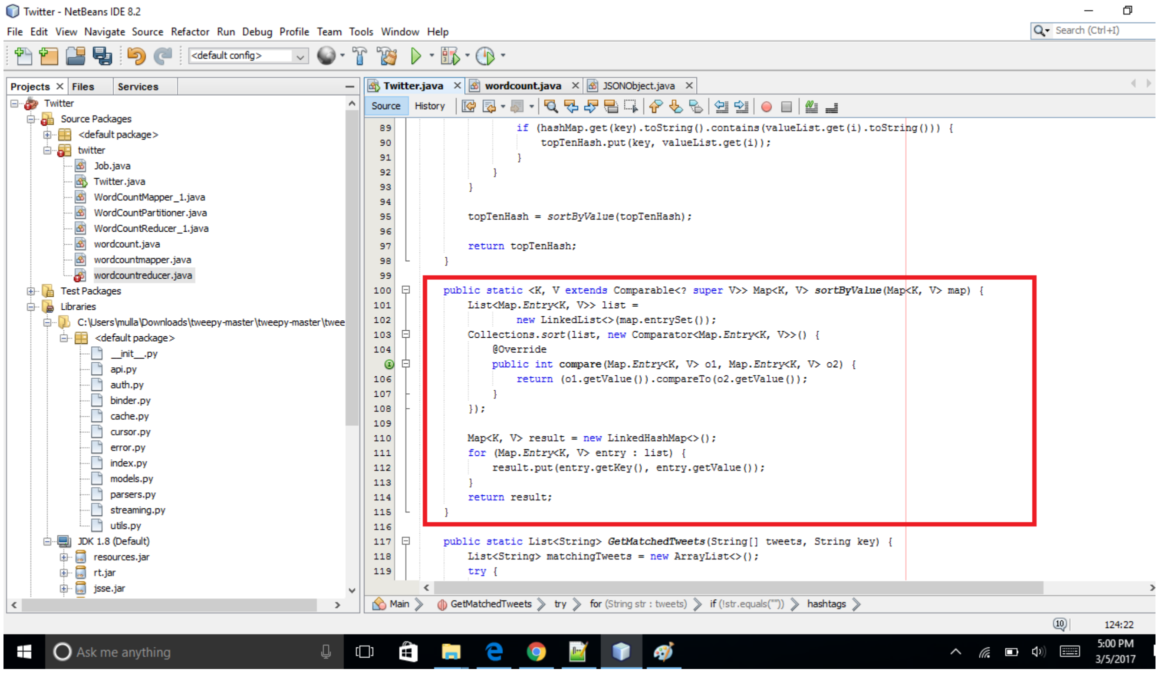
```
// contains the twitter data collection
package twitter
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class WordCountMapper extends
        Mapper<LongWritable, Text, Text, LongWritable> {
    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        // Read the line
        String line = value.toString();

        // Split the line into words
        String[] words = line.split(" ");

        // Assign count(1) to each word
        for (String word : words) {
            context.write(new Text(word), new LongWritable(1));
        }
    }
}
```

- Word Count Reduce Function Program



```java
// contains the twitter data collection
package twitter
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class WordCountReducer extends
        Reducer<Text, LongWritable, Text, LongWritable> {
    @Override
    protected void reduce(Text key, Iterable<LongWritable> values,
            Context context) throws IOException, InterruptedException {
        // Sum the List of values
        long sum = 0;
        for (LongWritable value : values) {
            sum = sum + value.get();
        }

        // Assign Sum to corresponding Word
        context.write(key, new LongWritable(sum));
    }

}
```

- Wordcount Sorting



```java
                if (hashMap.get(key).toString().contains(valueList.get(i).toString())) {
                    topTenHash.put(key, valueList.get(i));
                }
            }
        }

        topTenHash = sortByValue(topTenHash);

        return topTenHash;
    }

    public static <K, V extends Comparable<? super V>> Map<K, V> sortByValue(Map<K, V> map) {
        List<Map.Entry<K, V>> list =
                new LinkedList<>(map.entrySet());
        Collections.sort(list, new Comparator<Map.Entry<K, V>>() {
            @Override
            public int compare(Map.Entry<K, V> o1, Map.Entry<K, V> o2) {
                return (o1.getValue()).compareTo(o2.getValue());
            }
        });

        Map<K, V> result = new LinkedHashMap<>();
        for (Map.Entry<K, V> entry : list) {
            result.put(entry.getKey(), entry.getValue());
        }
        return result;
    }

    public static List<String> GetMatchedTweets(String[] tweets, String key) {
        List<String> matchingTweets = new ArrayList<>();
        try {
```