Principles og Big data -Twitter Project Phase:1

Team

- Balachandar Kulala(bkkhf)
- Ashish Motanam(amkhz)
- Ranjith (rvpbf)

Objective: "word counts of the extracted Hashtags and urls from the collected tweets using apache Hadoop and apache spark"

Tools Used: Hadoop, Apache Spark and Python.

Output:

The source code, log files and output files are uploaded into the following GitHub link.

GitHub Link: https://github.com/bkkhf/CS5540_PBDProject

Procedure:

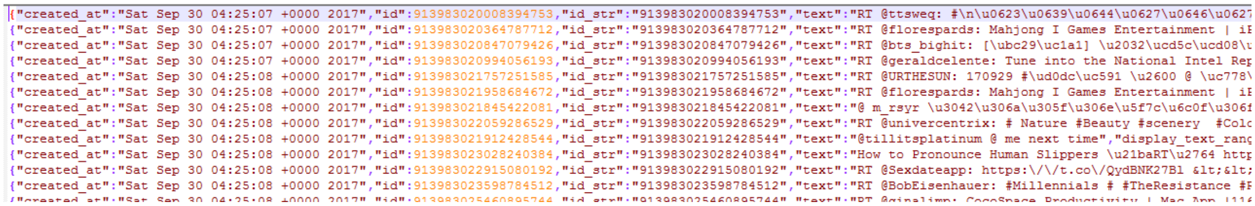
Step1: Collect the Tweets from Twitter using twitter API.

Collected the tweets using "tweepy.py" python module by using "#" as the filter.

The corresponding code is uploaded into the GitHub.

Source Code Name: "tweetStream.py"

Sample Collected Tweets: Uploaded the few tweets in the GitHub because of larger size.



```
{"created_at":"Sat Sep 30 04:25:07 +0000 2017","id":913983020008394753,"id_str":"913983020008394753","text":"RT @ttsweq: #\n\u0623\u0639\u0644\u0627\u0646\u0627\u0627
{"created_at":"Sat Sep 30 04:25:07 +0000 2017","id":913983020364787712,"id_str":"913983020364787712","text":"RT @florespards: Mahjong I Games Entertainment | iP
{"created_at":"Sat Sep 30 04:25:07 +0000 2017","id":913983020847079426,"id_str":"913983020847079426","text":"RT @bts_bighit: [\ubc29\uc1a1] \u2032\ucd5c\ucd08\u
{"created_at":"Sat Sep 30 04:25:07 +0000 2017","id":913983020994056193,"id_str":"913983020994056193","text":"RT @geraldcelente: Tune into the National Intel Rep
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983021757251585,"id_str":"913983021757251585","text":"RT @URTHESUN: 170929 #\ud0dc\uc591 \u2600 @ \uc778\
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983021958684672,"id_str":"913983021958684672","text":"RT @florespards: Mahjong I Games Entertainment | iP
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983021845422081,"id_str":"913983021845422081","text":"@ m_rsyr \u3042\u306a\u305f\u306e\u5f7c\u6c0f\u306f
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983022059286529,"id_str":"913983022059286529","text":"RT @univercentrix: # Nature #Beauty #scenery  #Colo
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983021912428544,"id_str":"913983021912428544","text":"@tillitsplatinum @ me next time","display_text_rang
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983023028240384,"id_str":"913983023028240384","text":"How to Pronounce Human Slippers \u21baRT\u2764 http
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983022915080192,"id_str":"913983022915080192","text":"RT @Sexdateapp: https:\/\/t.co\/QydBNK27Bl &lt;&lt;
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983023598784512,"id_str":"913983023598784512","text":"RT @BobEisenhauer: #Millennials # #TheResistance #P
{"created_at":"Sat Sep 30 04:25:08 +0000 2017","id":913983025460895744,"id_str":"913983025460895744","text":"RT @ginalimp: CocoSpace Productivity | Mac App |\u116
```

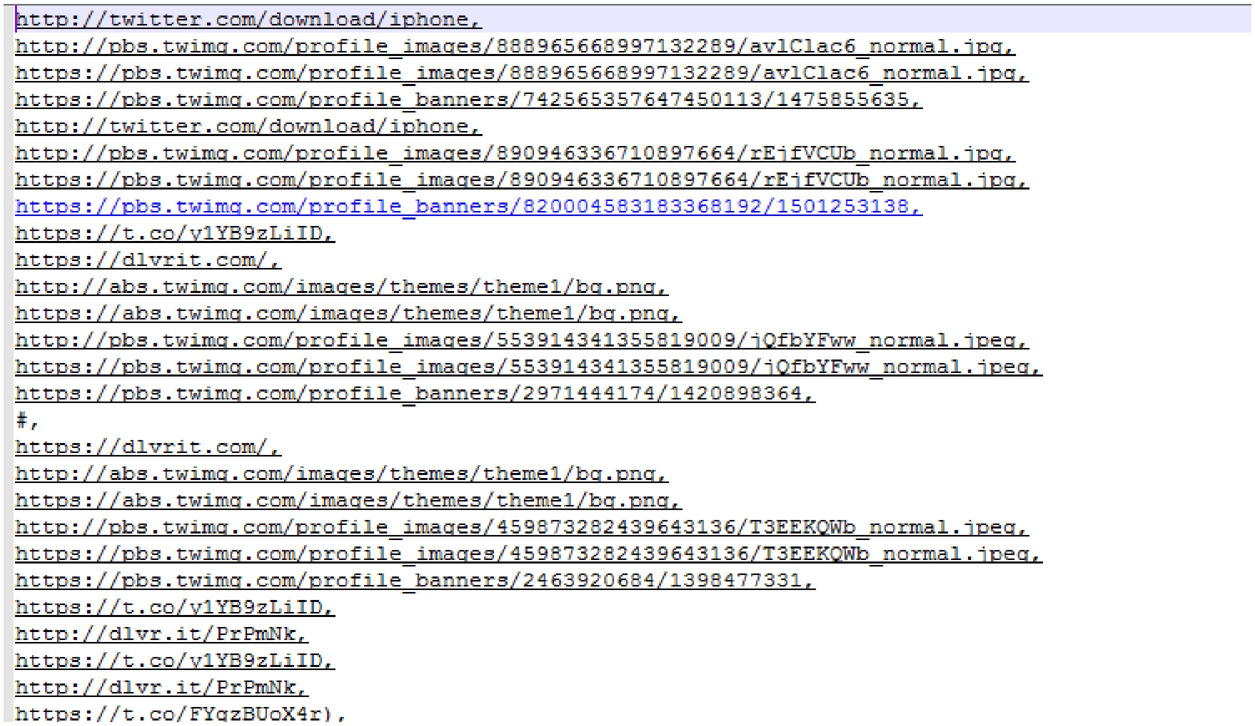
Step2: Extract the "Hashtags and URLs" from the collected tweets.

Extracts the HashTags and URLs from the collected tweets from the above step.

Source Code in the GitHub: "extractTweet.py"

Extracted Tweets are uploaded into GitHub: extractedHashTagURLs.txt

Sample Output:



```
http://twitter.com/download/iphone,
http://pbs.twimg.com/profile_images/888965668997132289/avlClac6_normal.jpg,
https://pbs.twimg.com/profile_images/888965668997132289/avlClac6_normal.jpg,
https://pbs.twimg.com/profile_banners/742565357647450113/1475855635,
http://twitter.com/download/iphone,
http://pbs.twimg.com/profile_images/890946336710897664/rEjfVCUb_normal.jpg,
https://pbs.twimg.com/profile_images/890946336710897664/rEjfVCUb_normal.jpg,
https://pbs.twimg.com/profile_banners/820004583183368192/1501253138,
https://t.co/y1YB9zLiID,
https://dlvrit.com/,
http://abs.twimg.com/images/themes/theme1/bg.png,
https://abs.twimg.com/images/themes/theme1/bg.png,
http://pbs.twimg.com/profile_images/553914341355819009/jQfbYFww_normal.jpeg,
https://pbs.twimg.com/profile_images/553914341355819009/jQfbYFww_normal.jpeg,
https://pbs.twimg.com/profile_banners/2971444174/1420898364,
#,
https://dlvrit.com/,
http://abs.twimg.com/images/themes/theme1/bg.png,
https://abs.twimg.com/images/themes/theme1/bg.png,
http://pbs.twimg.com/profile_images/459873282439643136/T3EEKQWb_normal.jpeg,
https://pbs.twimg.com/profile_images/459873282439643136/T3EEKQWb_normal.jpeg,
https://pbs.twimg.com/profile_banners/2463920684/1398477331,
https://t.co/y1YB9zLiID,
http://dlvr.it/PrPmNk,
https://t.co/y1YB9zLiID,
http://dlvr.it/PrPmNk,
https://t.co/FYqzBUoX4r),
```

Step3: Run word count example in apache Hadoop to get word count of hashtags and urls.

Step4: Run word count example in apache Spark to get word count of hashtags and urls.

Command to load the text file:

```
scala> val lines =
sc.textFile("/home/student/Installations/TwitterProject/extractedHashTagURLs.txt")
```

Command to split and count the values:

```
scala> val count=lines.flatMap(_.split(" ")).map(word => (word,1)).reduceByKey(_+_)
```

Command to write into a file:

```
scala> tools.nsc.io.File("/home/student/bala.txt").writeAll(count.collect().mkString(","))
```