Map – Reduced Based Twitter Analysis

get_tweets.py

Role – To Fetch Tweets from Twitter

- Get tweets from the twitter API using tweepy python library.
- Download Tweets with #Tag value given in the file
  - For Ex :- Download all tweets related to “trump” or “modi”
- Retrieve all tweets(in JSON format) & save into file.

Technical Section

- Download tweepy library.
- Register with twitter API get all below parameter.
- Required to give
  - access_token
  - access_token_secret
  - consumer_key
  - consumer_secret
- self.limit = How many tweets you want to Download.

filter_text.py

Role – To Extract “Tweets Text” from Downloaded Tweets.

Technical Section

- Open twitter file in “utf-8” encoding mode.
- Iterate over all tweets (JSON) & extract “text” property from it.
  - Ex : - { “ID” : 3453 , “time” : 12:30:45 , “text” : “car crashed”}
  - extract “car crashed” from it.
- Write all Tweets text into a new file
- This new file act as a input for mapper file.

mapper.py

Role – To make <key , value> pair & give it to reducer.

- For Example Your Input file contains three tweets
- “Car is stolen near new york street no.1” , “Tesla Model 3 launced” , “car summit 2016 held at perth”.
- The <key , value1 , value2 > is

<1,“Car is stolen near new york street no.1”,“Tesla Model 3 launced”>
<1 , “Car is stolen near new york street no.1”,“car summit 2016 held at perth”>
<2, “Tesla Model 3 launced”,“car summit 2016 held at perth”>

It make sense ….compare all tweets with each other…

Technical Section

- open the input file in “utf-8” mode.
- Iterate over two loops mechanism to to get <key ,value1 , value2> pairs.
- Key is the value of outer loops.
- It emits all <key,value> pairs so that reducer can process that <key,value> pair.

reduce.py

Role :- Give 90% similar tweets.

- Compare tweets , if they are 90 % similar to each other(character match) , it will emit to output file.

Technical Section

- Read all <key , value> pairs..
- For Example <key , value1 , value2> ..
- It will give <value1 ,value2> to difflib.SequenceMatcher …
- difflib.SequenceMatcher compare strings & give similarities of string in range of 0.0-1.0.
- if range >= 0.9 , it will emit that tweets into output file.

Hadoop Cluster

- Hadoop-2.7.3
- Run Map-Reduce job into hadoop (command)

```
/usr/local/hadoop/bin/hadoop jar
/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar \
-input /user/hduser/New/New/filtertext.txt \
-output /user/hduser/New/New/output \
-file  /home/hduser/final/mapper.py \
-file /home/hduser/final/reduce.py \
-mapper  /home/hduser/final/mapper.py \
-reducer /home/hduser/final/reduce.py
```