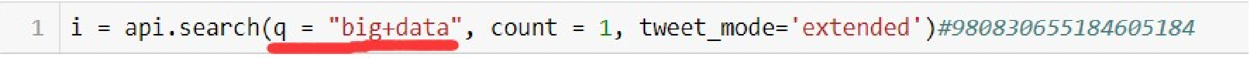
Lab 2 Part 2

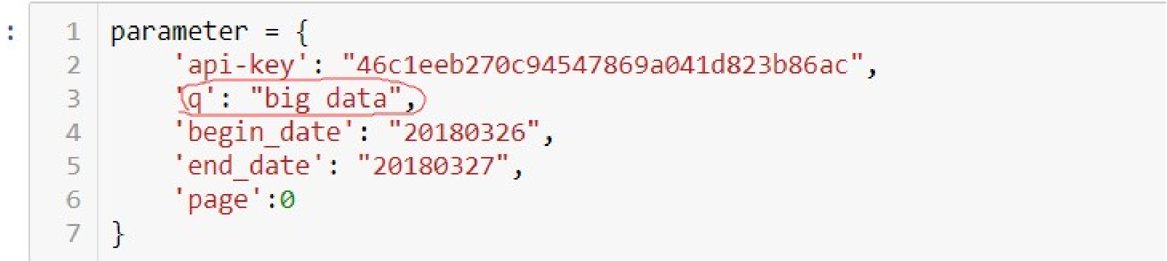
1. Get the data from Twitter with the notebook “tweets”. If you want to change another key word, just change the parameter of “q”.



```
i = api.search(q = "big+data", count = 1, tweet_mode='extended')#980830655184605184
```

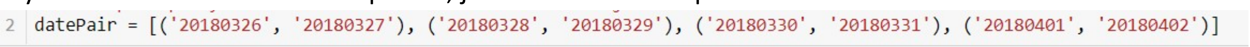
   Because the limit date of twitter is seven days, so the time period is 7 days, you can run this code until it throws some error.

2. Get the data from NYTimes. If you want to change another key word, just change the parameter “q”.



```
parameter = {
    'api-key': "46c1eeb270c94547869a041d823b86ac",
    'q': "big data",
    'begin_date': "20180326",
    'end_date': "20180327",
    'page':0
}
```

   If you want to extend the date period, just add some date pair of this:

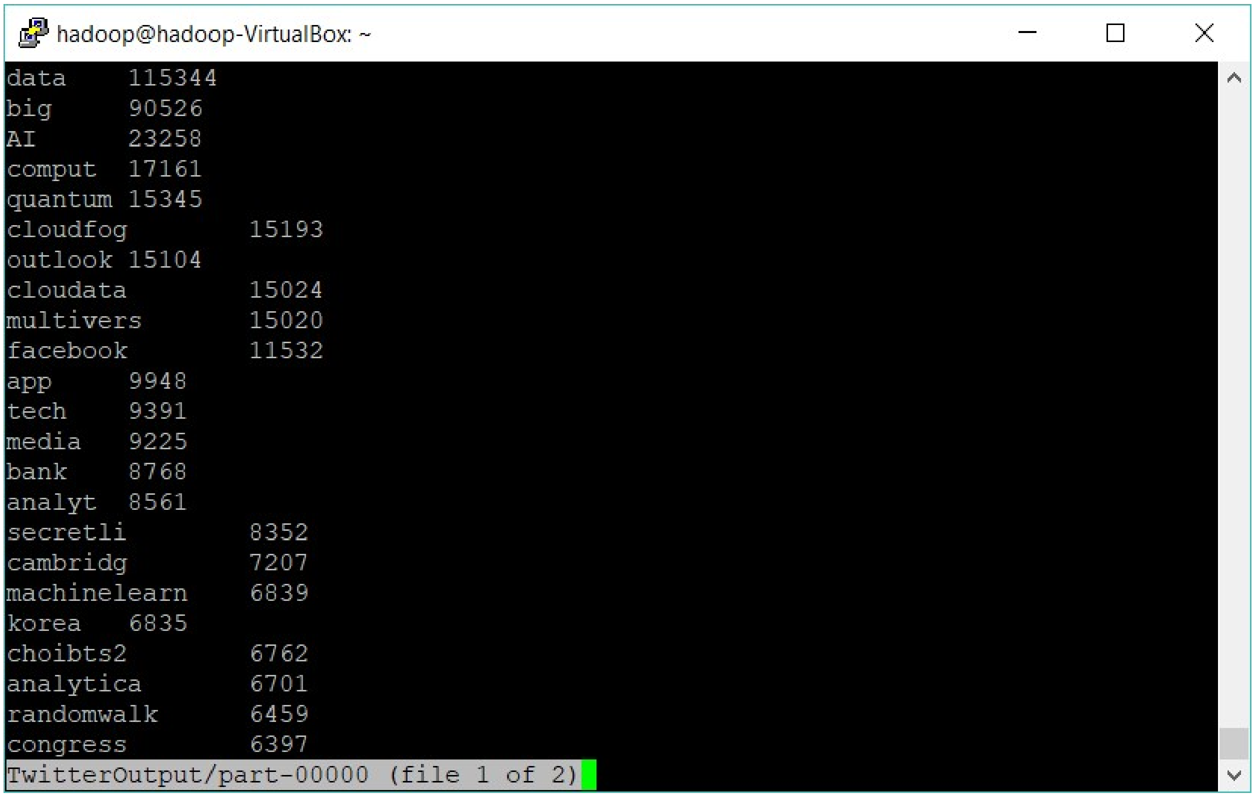


```
datePair = [('20180326', '20180327'), ('20180328', '20180329'), ('20180330', '20180331'), ('20180401', '20180402')]
```

   The code will get the data automatically.

3. After getting these data, move them into the VM and use the command ”hdfs dfs –put dir dir” to put the files into hdfs.

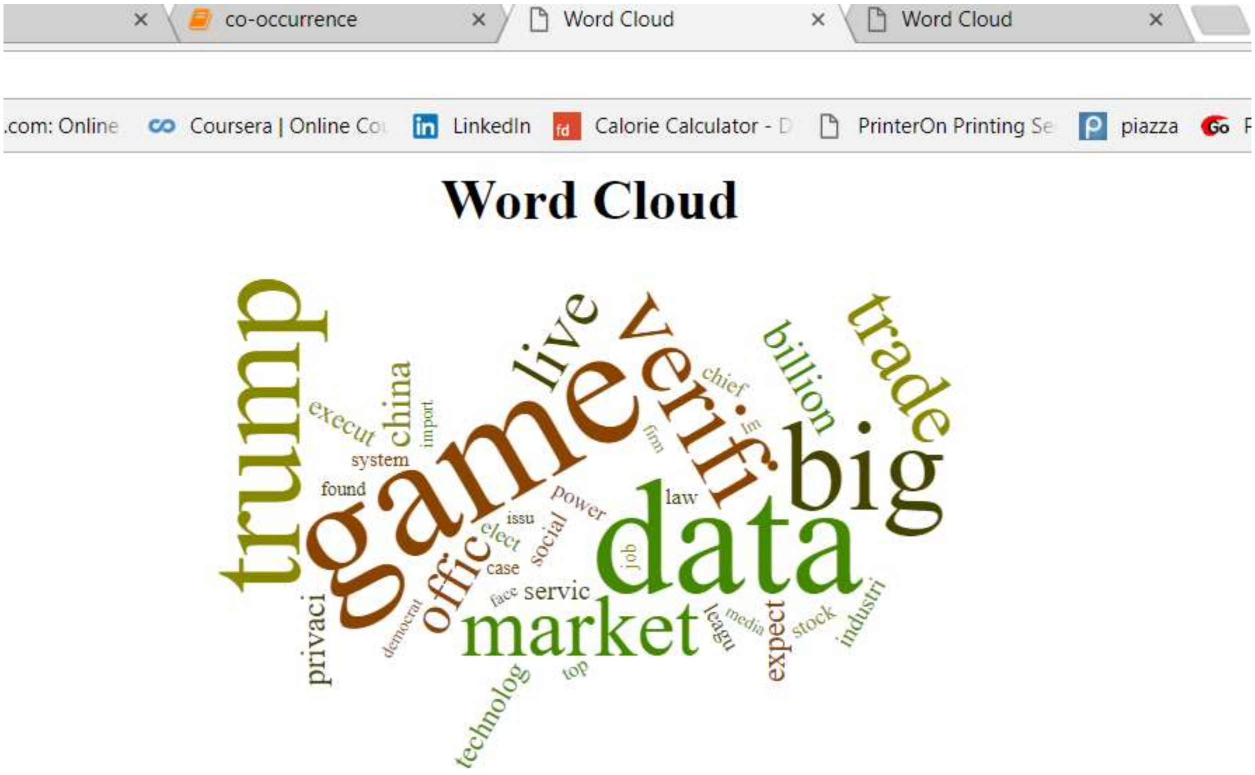
4. Use the files mapper.py and reducer.py to get the key pairs like this:



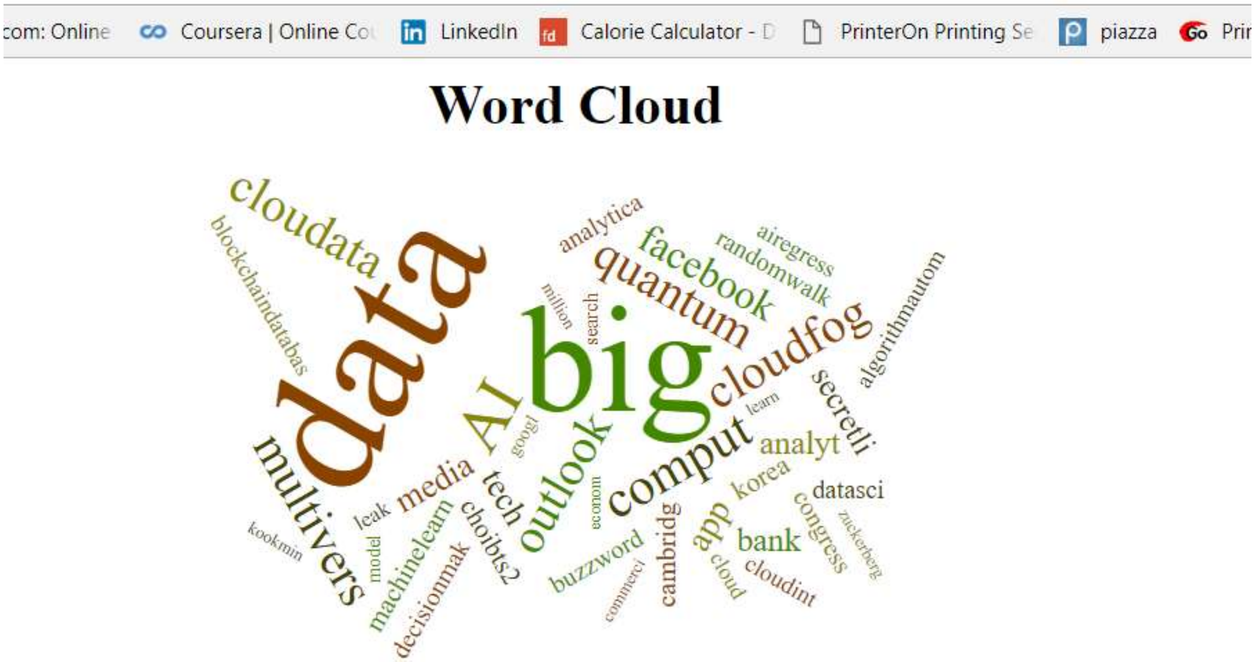
```
data    115344
big     90526
AI      23258
comput  17161
quantum 15345
cloudfog        15193
outlook 15104
cloudata        15024
multivers       15020
facebook        11532
app     9948
tech    9391
media   9225
bank    8768
analyt  8561
secretli        8352
cambridg        7207
machinelearn    6839
korea   6835
choibts2        6762
analytica       6701
randomwalk      6459
congress        6397
TwitterOutput/part-00000 (file 1 of 2)
```

For map-reduce there is another file is necessary which is stopwords.txt.

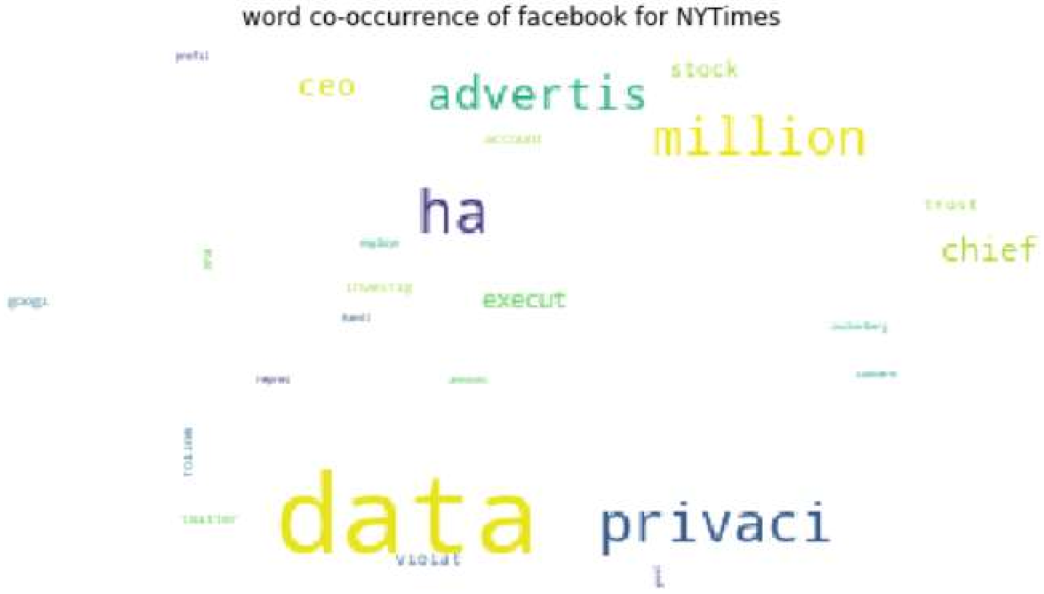
5. Use the notebook “wordcloud” to get a js file with key pairs we got from the last step.
6. Open the html in wordcloudNews\example to get the wordcloud page of NYTimes.



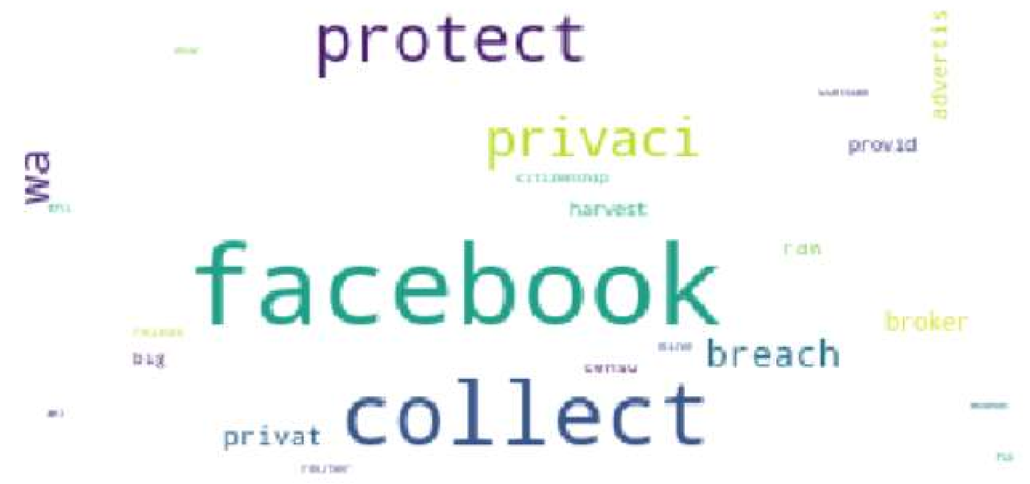
7. Open the html in wordcloudTwitter\example to get the wordcloud page of Twitter



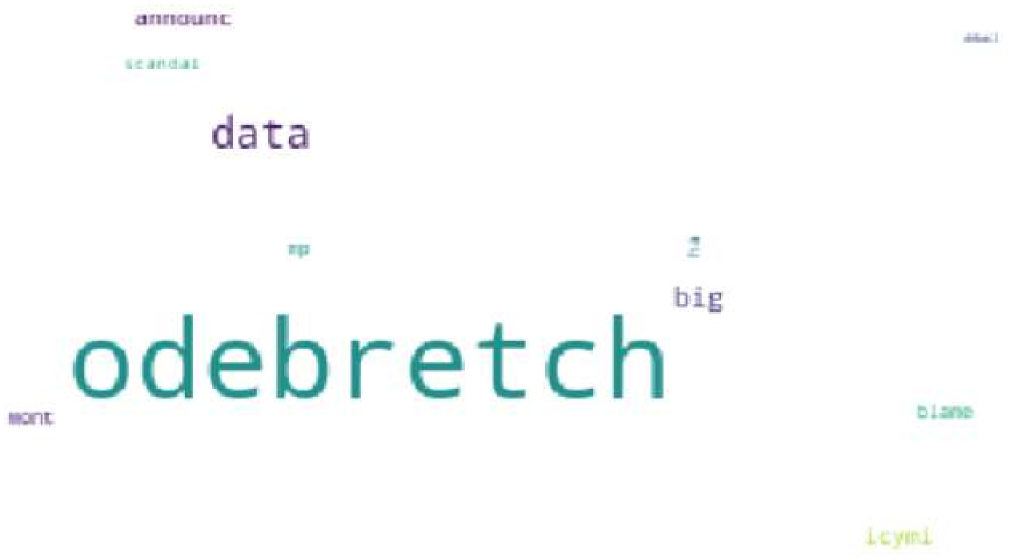
8. User the files coMapper.py, coReducer.py, and stopwords.txt to get co-occurrence of each data set.
9. To visualize the word co-occurrence we got from the last step, we can use the notebook “co-occurrence”, to generate some word cloud plot with python. The output is like this:



word co-occurrence of data for NYTimes



word co-occurrence of facebook for Twitter



Not all the results are given here in the report.