BIG DATA HOME ASSIGNMENT - 1

Dataset : Twitter Data

Topic : Pre-processing of twitter data

Mapper.py

```
#!/usr/bin/env python

from datetime import datetime

import json

import sys


# Mapper - filters tweets by user screen_name and generates key-value pairs of the format <hour_of_creation, 1>

filter_user_name = 'prezono'

for line in sys.stdin:

   tweet_data = json.loads(line)

   if tweet_data.get('user').get('screen_name').strip().lower() == filter_user_name:

      timestamp = tweet_data.get('created_at')

      date_object = datetime.strptime(timestamp, '%a %b %d %H:%M:%S +0000 %Y')

      print '%s\t%s' % (date_object.hour, 1)
```

Reducer.py

```
#!/usr/bin/env python


from __future__ import division

import sys

import string


# Reducer - processes the output of the mapper to compute the hourly average of tweets

tweet_count = 0
```

```
hour = None
tweets_by_hour_avg = {}
num_days = 365


for line in sys.stdin:
   (key, val) = line.strip().split('\t', 1)
   if hour != key:
      if hour:
         print "Average # of tweets @ %s hrs:\t%s" % (hour, tweet_count/num_days)
         tweets_by_hour_avg[hour] = tweet_count/num_days
         tweet_count = 0
   hour = key
   try:
      tweet_count += int(val)
   except:
      continue
print 'Tweets @ %s hrs:\t%s' % (hour, tweet_count)
tweets_by_hour_avg[hour] = tweet_count/num_days


print "Hour of the day that @PrezOno tweets the most:", max(tweets_by_hour_avg,
key=tweets_by_hour_avg.get)
```

Output :

sample

{created_at:Tue Feb 11 04:45:51 +0000 2014,id_str:433099539201806336}

```
{"created_at":"Tue Feb 11 04:45:51 +0000 2014","id":433099539201806336,"id_str":"433099539201806336"
{"created_at":"Tue Feb 11 04:45:56 +0000 2014","id":433099562358931456,"id_str":"433099562358931456"
{"created_at":"Tue Feb 11 04:46:00 +0000 2014","id":433099579056480256,"id_str":"433099579056480256"
{"created_at":"Tue Feb 11 04:46:02 +0000 2014","id":433099584902938624,"id_str":"433099584902938624"
{"created_at":"Tue Feb 11 04:46:02 +0000 2014","id":433099587067195392,"id_str":"433099587067195392"
{"created_at":"Tue Feb 11 04:46:04 +0000 2014","id":433099594377883649,"id_str":"433099594377883649"
{"created_at":"Tue Feb 11 04:46:05 +0000 2014","id":433099598379225088,"id_str":"433099598379225088"
{"created_at":"Tue Feb 11 04:46:07 +0000 2014","id":433099605023412224,"id_str":"433099605023412224"
{"created_at":"Tue Feb 11 04:46:07 +0000 2014","id":433099604872019968,"id_str":"433099604872019968"
{"created_at":"Tue Feb 11 04:46:09 +0000 2014","id":433099616507428864,"id_str":"433099616507428864"
{"created_at":"Tue Feb 11 04:46:12 +0000 2014","id":433099626405580801,"id_str":"433099626405580801"
{"created_at":"Tue Feb 11 04:46:14 +0000 2014","id":433099637613137920,"id_str":"433099637613137920"
{"created_at":"Tue Feb 11 04:46:16 +0000 2014","id":433099644193624065,"id_str":"433099644193624065"
{"created_at":"Tue Feb 11 04:46:20 +0000 2014","id":433099663021862913,"id_str":"433099663021862913"
{"created_at":"Tue Feb 11 04:46:21 +0000 2014","id":433099666234671104,"id_str":"433099666234671104"
{"created_at":"Tue Feb 11 04:46:24 +0000 2014","id":433099678025261056,"id_str":"433099678025261056"
{"created_at":"Tue Feb 11 04:46:31 +0000 2014","id":433099705443430400,"id_str":"433099705443430400"
{"created_at":"Tue Feb 11 04:46:31 +0000 2014","id":433099705388503040,"id_str":"433099705388503040"
{"created_at":"Tue Feb 11 04:46:33 +0000 2014","id":433099715073564672,"id_str":"433099715073564672"
{"created_at":"Tue Feb 11 04:46:38 +0000 2014","id":433099736816844801,"id_str":"433099736816844801"
{"created_at":"Tue Feb 11 04:46:39 +0000 2014","id":433099739601448960,"id_str":"433099739601448960"
{"created_at":"Tue Feb 11 04:46:43 +0000 2014","id":433099756127387649,"id_str":"433099756127387649"
{"created_at":"Tue Feb 11 04:46:48 +0000 2014","id":433099777165631488,"id_str":"433099777165631488"
{"created_at":"Tue Feb 11 04:46:49 +0000 2014","id":433099781419048960,"id_str":"433099781419048960"
```

BIG DATA HOME ASSIGNMENT - 2

Dataset : Twitter Data

Topic : Analysing Pre-processed twitter data

Mapper.py

```
#!/usr/bin/env python
from datetime import datetime
import json
import sys

# Mapper - filters tweets by user screen_name and generates key-value pairs of the format <hour_of_creation, 1>
filter_user_name = 'prezono'
for line in sys.stdin:
   tweet_data = json.loads(line)
   if tweet_data.get('user').get('screen_name').strip().lower() == filter_user_name:
      timestamp = tweet_data.get('created_at')
      date_object = datetime.strptime(timestamp, '%a %b %d %H:%M:%S +0000 %Y')
      print '%s\t%s' % (date_object.hour, 1)
```

Reducer.py

```
#!/usr/bin/env python

from __future__ import division
import sys
import string

# Reducer - processes the output of the mapper to compute the hourly average of tweets
tweet_count = 0
```

```
hour = None
tweets_by_hour_avg = {}
num_days = 365


for line in sys.stdin:
   (key, val) = line.strip().split('\t', 1)
   if hour != key:
      if hour:
         print "Average # of tweets @ %s hrs:\t%s" % (hour, tweet_count/num_days)
         tweets_by_hour_avg[hour] = tweet_count/num_days
         tweet_count = 0
   hour = key
   try:
      tweet_count += int(val)
   except:
      continue
print 'Tweets @ %s hrs:\t%s' % (hour, tweet_count)
tweets_by_hour_avg[hour] = tweet_count/num_days


print "Hour of the day that @PrezOno tweets the most:", max(tweets_by_hour_avg,
key=tweets_by_hour_avg.get)
```

Output:

Sample:

```
Average # of tweets @ 0 hrs:    0.0383561643836
Average # of tweets @ 1 hrs:    0.0520547945205
Average # of tweets @ 10 hrs:   0.0438356164384
Average # of tweets @ 11 hrs:   0.0657534246575
Average # of tweets @ 12 hrs:   0.0328767123288
Average # of tweets @ 13 hrs:   0.041095890411
```

```
Average # of tweets @ 0 hrs:    0.0383561643836
Average # of tweets @ 1 hrs:    0.0520547945205
Average # of tweets @ 10 hrs:   0.0438356164384
Average # of tweets @ 11 hrs:   0.0657534246575
Average # of tweets @ 12 hrs:   0.0328767123288
Average # of tweets @ 13 hrs:   0.041095890411
Average # of tweets @ 14 hrs:   0.0547945205479
Average # of tweets @ 15 hrs:   0.0356164383562
Average # of tweets @ 16 hrs:   0.0246575342466
Average # of tweets @ 17 hrs:   0.0712328767123
Average # of tweets @ 18 hrs:   0.027397260274
Average # of tweets @ 19 hrs:   0.0520547945205
Average # of tweets @ 2 hrs:    0.0438356164384
Average # of tweets @ 20 hrs:   0.0575342465753
Average # of tweets @ 21 hrs:   0.0356164383562
Average # of tweets @ 22 hrs:   0.0520547945205
Average # of tweets @ 23 hrs:   0.0438356164384
Average # of tweets @ 3 hrs:    0.0575342465753
Average # of tweets @ 4 hrs:    0.0438356164384
Average # of tweets @ 5 hrs:    0.00821917808219
Average # of tweets @ 6 hrs:    0.0027397260274
Average # of tweets @ 7 hrs:    0.0109589041096
Average # of tweets @ 8 hrs:    0.00821917808219
Average # of tweets @ 9 hrs:    0.03013698630136
Hour of the day that @PrezOno tweets the most: 17
```