**P4 & P5: *Hadoop for Real Problems***

Group Members: Surbhi Bhatnagar(bhatnash), Meenakshi Venkat(venkatmi), Ryan Parker(parkerr3), Xiaoting Zhu(zhuxt)

Twitter Dataset Questions done: 1,2,3,5,6,7,10

**1. What hour of the day does @PrezOno’s tweet the most on average, using every day we have twitter data? Include a plot of the expected number of tweets for each hour of the day, for those he did tweet.**

For example if PrezOno tweeted once every day at 12:30PM, his expected number of tweets between 12 and 1 would be 1. If he alternates between 2 and 3 tweets per day, his average would be 2.5.

Answer1:

The following commands in PigLatin to obtain results for question 1:

**tweet\_raw** = load '/data/twitter/' as (raw: chararray);

**tweets** =foreach tweet\_raw generate REGEX\_EXTRACT(raw, 'created\_at":"([^"]+)"',1) as created\_at,REGEX\_EXTRACT(raw, 'created\_at":"([^ ]+) ',1) as wday,REGEX\_EXTRACT(raw, ' (\\d\\d):\\d\\d:\\d\\d',1) as hday, REGEX\_EXTRACT(raw, '"text":"([^"]+)"',1) as text, REGEX\_EXTRACT(raw, '"id\_str":"(\\d+)","name"',1) as id;

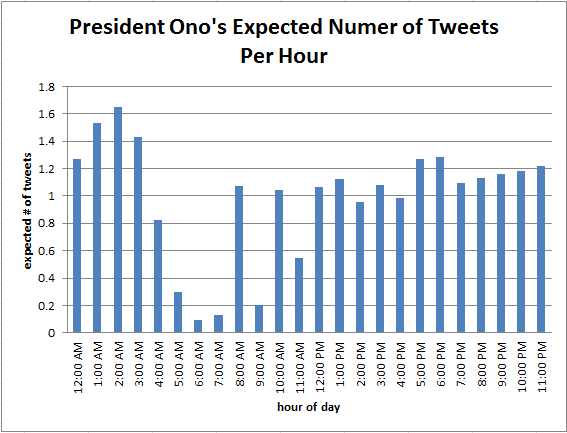
**prezono** = filter tweets by id=='211178363';

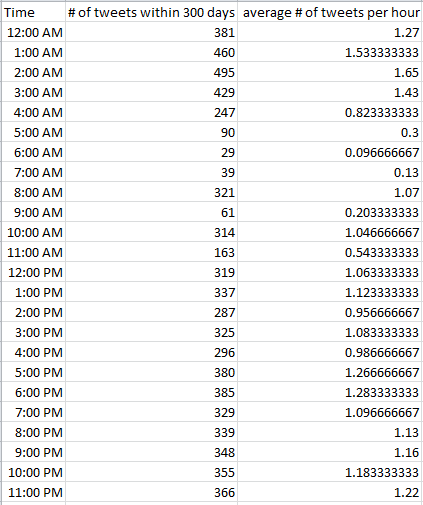
**tweet\_count** = foreach prezono generate hday;

**hour\_count** = group tweet\_count by hday parallel 24;

**hour\_count2** = foreach hour\_count generate $0 as hour,COUNT(tweet\_count) as count;

The results of the above sequence is displayed below. This data shows that President Ono is most likely to tweet between 2:00 a.m. and 3:00 a.m.





**2. What day of the week does @PrezOno tweet the most on average? Use the same example as in #1 but for days of the week.**

Answer 2: Wednesday

Using Hadoop in streaming mode, a mapreduce job was applied to the twitter data of February 2015. The mapper split and reads each tweet as one line. For each tweet the screen name and day/date contents were found. It checks the screen name for President Ono’s to distinguish which tweets were his and would increment a count for the number of his tweets as well as store the day of the week it was tweeted. The mapper then send the summary for each day of the week containing the tweet count of PrezOno on each of the days. The reducer then combines data from all mappers and finds total count of PrezOno tweets for each day of the week. It is then printed along with the average calculated by dividing with total number of PrezOno Tweets.

Results for partial output file submitted on github along with our code:

Total tweets of President Ono: 241

Total and Average Tweets for monday: 51 0.211618

Total and Average Tweets for tuesday: 21 0.087137

Total and Average Tweets for wednesday: 64 0.265560

Total and Average Tweets for thursday: 14 0.058091

Total and Average Tweets for friday: 17 0.070539

Total and Average Tweets for saturday: 39 0.161826

Total and Average Tweets for sunday: 35 0.145228

For entire twitter data:

Total tweets of President Ono: 7095

Total and Average Tweets for monday: 838 0.118111

Total and Average Tweets for tuesday: 962 0.135588

Total and Average Tweets for wednesday: 1090 0.153629

Total and Average Tweets for thursday: 1113 0.156871

Total and Average Tweets for friday: 1064 0.149965

Total and Average Tweets for saturday: 969 0.136575

Total and Average Tweets for sunday: 1059 0.149260

**3. How does @PrezOno’s tweet length compare to the average of all others? What is his average length? All others?**

Answer 3:

Using Hadoop in streaming mode, a mapreduce job was applied to the twitter data of February 2015.

The mapper split each tweet as one line. For each tweet the screen name and contents were found. It would check the screen name for President Ono’s to distinguish which tweets were his and would increment a count for the number of his tweets as well as determine the length of the content. A separate count was incremented by the content length. If the screen name was not President Ono’s the same functions previously stated would be applied to two other counts. So in the end the mapper sent a count for number of President Ono’s tweets, number of all other user tweets, sum of all the lengths of President Ono’s tweets, and the sum of the length of all other users’ tweets.

The reducer calculated the results shown below using the values provided by the mapper.

Average Tweet Length of President Ono: 102.987551867

Average Tweet Length of other users: 60.273566583

Total Tweets: 864996

Total PTweets: 241

**6. For each day of the week, what was the most mentioned hashtag? Hour of the day?**

Answer 6:

We calculated the most mentioned hashtag for each day of the week in the twitter data. We also calculated the most mentioned hashtag for each hour. For this purpose, we create a dictionary in the map function for both the hour and day of the week. We pass the day(hour) along with the hashtag and sum of the hashtags mentioned for that particular day(hour) to the reducer. In the reducer, we calculate the total for each hashtag per day and then find the maximum per day. We performed this analysis in stream mode using python.

Output of partial twitter data testing is as follows:

00 shakesbowl 15

01 justsayin 46

02 hahn 66

03 gobucks 9

04 swag 2

05 job 32

06 cincinnati 5

07 nursing 8

08 ohio 19

09 job 21

10 nursing 5

11 columbus 6

12 bearcats 244

13 cincinnati 19

14 columbus 58

15 new 15

16 gobucks 9

17 deliciousautomobiles 9

18 physician 8

19 thestruggle 2

20 changingprepaid 31

21 peopleyouseeinanairportterminal 14

22 coshocton 4

23 sb49 296

Fri snowweek 6

Mon superbowlchamps2015 52

Sun coshocton 4

Thu octavia2015 1

Tue hr 13

Wed autocheck 2

**7. For those tweets with location information, what lat/long (or city/state) is the centroid? What was the proportion of tweets with location to those without?**

Answer 7:

We calculated the longitude and latitude for each tweet that was present in the twitter data.Some data did not have these values but for the rest of the data we used the information available to calculate the average lat/long.

For calculating the average lat/long, we calculate the total long/lat and then total number of tweets with lat/long values and the total number of tweets without this information for each block in the map function. Later in the reducer, we sum of the values and find the average lat/long and the number of tweets without the information. We performed this analysis in stream mode using python.

Total tweets : 121684141

tweets with lat/long value:37367389

Tweets without lat/long values:84316752

Average latitude:38.9873

Average Longitude :-82.4416

0.69 percent of the entire tweets did not have location information in their tweets. And the tweets with the location information mainly were from Cincinnati.Thus, the centroid lat/long seems to be close to cincinnati(**Latitude:** 39.1619444 **Longitude:** -84.4569444).

**10. For each day, what was the most retweeted or most favorited tweet?**

Answer 10 :

We calculated the most retweeted or favorite tweet for each day present in the twitter dataset. In the map function, we pass only the top tweets for each day present in the block. Here the date is the key and the text and retweet count are merged and sent to the reducer. In the reducer, we append all the values for each date and then print only the maximum retweet value and tweet. We get maximum favorited/retweeted tweet for every day. A Part of the output file is as follows:

Feb20 Congrats to @Lawdagger23 for being named Birmingham Tip Off Club's Player of the Year! #FAST32 #BearcatNation http://t.co/qGDUxUPuyu [58]

Feb21 The first of a pair today! Ryan DePietro wins the 400 IM at the @American\_Swim Champs. #Bearcats http://t.co/HsnR7Ww2zo [49]

Feb22 #Bearcats come back to claim 63-53 win at Houston. Thomas with a career-high 18. Johnson had 12 and Clark with 11 points &amp; 14 rebounds. [87]

Feb23 Best Animated Short: 'Feast'! Tri-Stater @PatrickTOsborne (Who went to St.X!) Just won an OSCAR for this! #Oscars #Cincyproud [131]

Feb19 Just announced: the #Bearcats 2015 Football Schedule! #Nippert2015 http://t.co/xs7d4TwYk5 [239]

Feb18 This shows who is better #HottestCollegeinAmerica http://t.co/7Yz9bShJh[16]

Jan21 Create your very own #Cincy in a Box and we'll ship it for free. http://t.co/9wal5KTolC [4]

Feb04 Check out the top feeder high schools to University of Cincinnati's Honors Programs http://t.co/fd5Dsc60xI [110]

**5. What twitter user tweeted the most? What is the longest tweeters? And Bottom?**

Answer 5:

This analysis was performed using streaming mode in python

Firstly the file was parsed into dictionary using json.load(). In order to find the user that tweeted the most, we calculated the tweeted time for every user. The user id [‘user’][‘id\_str’] was used as the key and the value increased by 1 when an id was scanned. To increase the efficient, a hash method was used to parse the output key into 0~10000. By dividing the user id by 10000, we made the output as [hashv, user\_ids, tweeted\_times]. Within each hash block that containing the same user\_ids, it was easier to calculate the tweeted times of each user\_id. The hash value and max tweeted times were returned, together with the user\_id. After scanning the max tweeted times for each hash block we could get the max tweeted time for all the dataset and the respective user\_id.

./id.sh /data/twitter/2015-02-22.txt output

For the longest and shortest tweeters, we summed the total twitter length for each user by introducing dictionary l{}. The average twitter length came from the total twitter length dividing by tweeted times. With the same hash method ranking the average twitter length, we could get the user\_id with the longest and shortest twitter length.

Results for the whole dataset:

The user tweeted most:

Max: [68637, '2433105217']

which means user '2433105217' tweeted 68637 times

The other questions were not able to be run in streaming mode, so partial dataset (2015-02-22.txt) was used to do the analysis

./long.sh /data/twitter/2015-02-22.txt longest

./short.sh /data/twitter/2015-02-22.txt shortest

Results on partial datasets:

Max: [300, '26607998']

which means user ‘26607998’ tweeted 300 times

Longest: [166, '430227252']

User ‘430227252’ had the longest average twitter length of 166 characters.

Shortest: [1, '2705588739']

User '2705588739' had the shortest average twitter length of 1 character.

Several other partial dataset running on local mode:

