**Cloud Computing P3 -** ***Hadoop For Real Problems***

Team Members :

Madhu babu Anchuri (M08740177)

Hemanth Vemulaplli (M08651035)

Vinod Challa (M08843643)

Sairam Doppalapudi (M08740353)

We have chosen the twitter data for this project and solved four questions. We have used streaming mode and wrote jobs in python.

Questions we have solved are :

Q3. How does @PrezOno’s tweet length compare to the average of all others?  What is his average length?  All others?

Q4. Create a 2d [heatmap](https://www.google.com/search?q=gnuplot+heat+map&oq=gnuplot+hea&aqs=chrome.1.69i57j0l5.3118j0j7&sourceid=chrome&espv=210&es_sm=119&ie=UTF-8) of tweet length (y axis, 7 or more bins) to hour of the day (x axis) for everyone.  You’ll have a grid of 168 colored boxes.  Be sure to normalize each box’s color.

Q5. What twitter user tweeted the most?  What is the top 5 longest tweeters over each’s average tweet length?  Bottom 5?

Q6. For each day of the week, what was the most mentioned hashtag?  Hour of the day?

Q3. How does @PrezOno’s tweet length compare to the average of all others?  What is his average length?  All others?

**Approach :**

First, Calculated PrezOno’s tweet length, count and others tweet length, count separately and then calculated the average of PrezOno’s and others seperately

**Procedure followed:**

* Read the data in streaming mode
* Wrote a getText function to return the following in list format

[‘prezOno’, [len(text),1]] and [‘others’,[len(text),1]]

* By using reducedByKey created rdd’s of tweet length and tweetcount.

Eg: texts.map(lambda a: (a[0],a[1][0])).reduceByKey(lambda a,b: (a+b))

* Then calculated the average of PrezOno and remaining others by using above tweetcount and tweetlength values.
* Saved the results to a file using saveAsTextFile.

**Final Result:**

Ono\_tweetavg 104

other\_tweetavg 81

Q4. Create a 2d [heatmap](https://www.google.com/search?q=gnuplot+heat+map&oq=gnuplot+hea&aqs=chrome.1.69i57j0l5.3118j0j7&sourceid=chrome&espv=210&es_sm=119&ie=UTF-8) of tweet length (y axis, 7 or more bins) to hour of the day (x axis) for everyone.  You’ll have a grid of 168 colored boxes.  Be sure to normalize each box’s color.

**Approach:**

Calculate tweet length for every hour seperately and then add them by hour key so that we get the tweet length per each and then plotted the graph.

**Procedure followed:**

* Read the data in the streaming mode
* Wrote a getText function which returns hour and length of the tweet.

Eg: (time[3].split(‘:’)[0],lengths)

* Using the reducedByKey calculated tweetlength for each hour.
* Output is coalesced into a single file using saveAsTextFile.

Eg: texts.coalesce(1).aveAsTextFile(“”)

**Result:**

(u'11', 11954841)

(u'10', 10068439)

(u'13', 17043586)

(u'12', 14364949)

(u'15', 22449524)

(u'14', 19879349)

(u'17', 24543109)

(u'16', 23820628)

(u'19', 25145996)

(u'18', 24580634)

(u'02', 28765627)

(u'03', 26706647)

(u'00', 25480610)

(u'01', 27723636)

(u'06', 13411509)

(u'07', 10952051)

(u'04', 22180969)

(u'05', 17404435)

(u'08', 9488530)

(u'09', 9064081)

(u'20', 25419439)

(u'21', 25516438)

(u'22', 24644165)

(u'23', 24812387)

**HeatMap:**

|  |  |
| --- | --- |
| hour | tweetlength |
| 0 | 25480610 |
| 1 | 27723636 |
| 2 | 28765627 |
| 3 | 26706647 |
| 4 | 22180969 |
| 5 | 17404435 |
| 6 | 13411509 |
| 7 | 10952051 |
| 8 | 9488530 |
| 9 | 9064081 |
| 10 | 10068439 |
| 11 | 11954841 |
| 12 | 14364949 |
| 13 | 17043586 |
| 14 | 19879349 |
| 15 | 22449524 |
| 16 | 23820628 |
| 17 | 24543109 |
| 18 | 24580634 |
| 19 | 25145996 |
| 20 | 25419439 |
| 21 | 25516438 |
| 22 | 24644165 |
| 23 | 24812387 |

Q5. What twitter user tweeted the most?  What is the top 5 longest tweeters over each’s average tweet length?  Bottom 5?

**Approach:**

For the first part, it is enough to calculate the tweetcount for every user and take the user who has the maximum count.

To Calculate the Top5 and Bottom5 first calculate the average tweet length and sort them. Then take the top5 and bottom5 from the sorted list.

**Procedure Followed :**

* Read the data in streaming mode
* Wrote getText function which returns name,,tweetlength and 1.

Eg: [name,(len(text),1)]

* Created map and applied reduceByKey on above returned values to sum tweetlength and count for each user.(rdd is created here)

Eg: tweets.map(getText).reduceByKey(lambda a,b: (a[0]+b[0],a[1]+b[1])

* And then Calculated the average for every user.(rdd is created here)
* By using the takeOrdered sorted the average and count in ascending and descending order separately.

Eg: avg.takeOrdered(5,key=lambda x: x[1])

* Lists are created above
* To save them to a textfile above lists are converted to rdd’s using sc.parellalize
* Saved the output’s to different files.

**Result:**

**Bottom 5:**

(u'Laila\_Lafrai', 0.0)

(u'Fun\_Size20', 0.0)

(u'Trevorsturgill5', 0.0)

(u'2013Afi9', 0.0)

(u'Im\_Lil\_Wanie', 0.0)

**Top 5:**

(u'Huntersweat', 416.0)

(u'RoyalEliteKiva', 350.0)

(u'blackxhole', 319.0)

(u'KelleeMichele', 272.0)

(u'pizzadellarry', 253.0)

**Most Tweeted User:**

(u'marilyn9743', 3419)

Q6. For each day of the week, what was the most mentioned hashtag?  Hour of the day?

**Approach:**

There are two parts in the question.Need to find most mentioned hashtag for each day of the week and also for each hour of the day separately.

For the first part, we need to find the hashtags for each day of the week and then reduce them using the weekday as a key and then find the hashtag which has maximum count on each day of the week.

For the second part, we need to find the hashtags for each hour of the day and then reduce them by using hour of the day and then finding the hashtag which has maximum count for each hour.

**Procedure Followed:**

* Read the data in streaming mode
* Wrote three functions getText, gethour, hashcount.
* getText returns the day and the hashtags mentioned of every tweet .
* gethour returns the hour and the hashtags mentioned of every tweet.
* Hashcount takes the output from above two functions as input and calculates the hashcount which is maximum for each hour and each day of the week separately.
* Then the above results are combined and written to a single file using saveAsText.

**Result:**

**Most Mentioned Hashtag for each day of the week :**

(u'Sat', u'dog', 26930)

(u'Tue', u'dog', 29171)

(u'Sun', u'dog', 29388)

(u'Fri', u'dog', 27051)

(u'Mon', u'dog', 28820)

(u'Wed', u'dog', 28264)

(u'Thu', u'dog', 26295)

**Most Mentioned hashtag for each hour of the day:**

(u'11', u'dog', 7814)

(u'10', u'dog', 7101)

(u'13', u'dog', 9042)

(u'12', u'dog', 8598)

(u'15', u'dog', 9861)

(u'14', u'dog', 9594)

(u'17', u'dog', 9328)

(u'16', u'dog', 9503)

(u'19', u'dog', 9488)

(u'18', u'dog', 9145)

(u'02', u'winning', 8185)

(u'03', u'winning', 8066)

(u'00', u'dog', 8218)

(u'01', u'dog', 8485)

(u'06', u'dog', 5850)

(u'07', u'dog', 5998)

(u'04', u'dog', 6803)

(u'05', u'dog', 6244)

(u'08', u'dog', 6158)

(u'09', u'dog', 6613)

(u'20', u'dog', 9156)

(u'21', u'dog', 9317)

(u'22', u'dog', 8647)

(u'23', u'dog', 8898)