**Project P3 Report**

Team:

Mallepalli Rakesh Reddy

Kakumanu Satya Harshavardhan

Sai Pradeep Chokka Sudarsan

Sreekhar Ale

Google 1gram questions (hdfs://data/1gram  [link](http://storage.googleapis.com/books/ngrams/books/datasetsv2.html)):

1. For each year available, plot the size of the set of words used.  Year on the x-axis, number of words on y-axis.

**Ans:**

This question requires number of words used in each year from 1505 to 2008.

**Approach:**

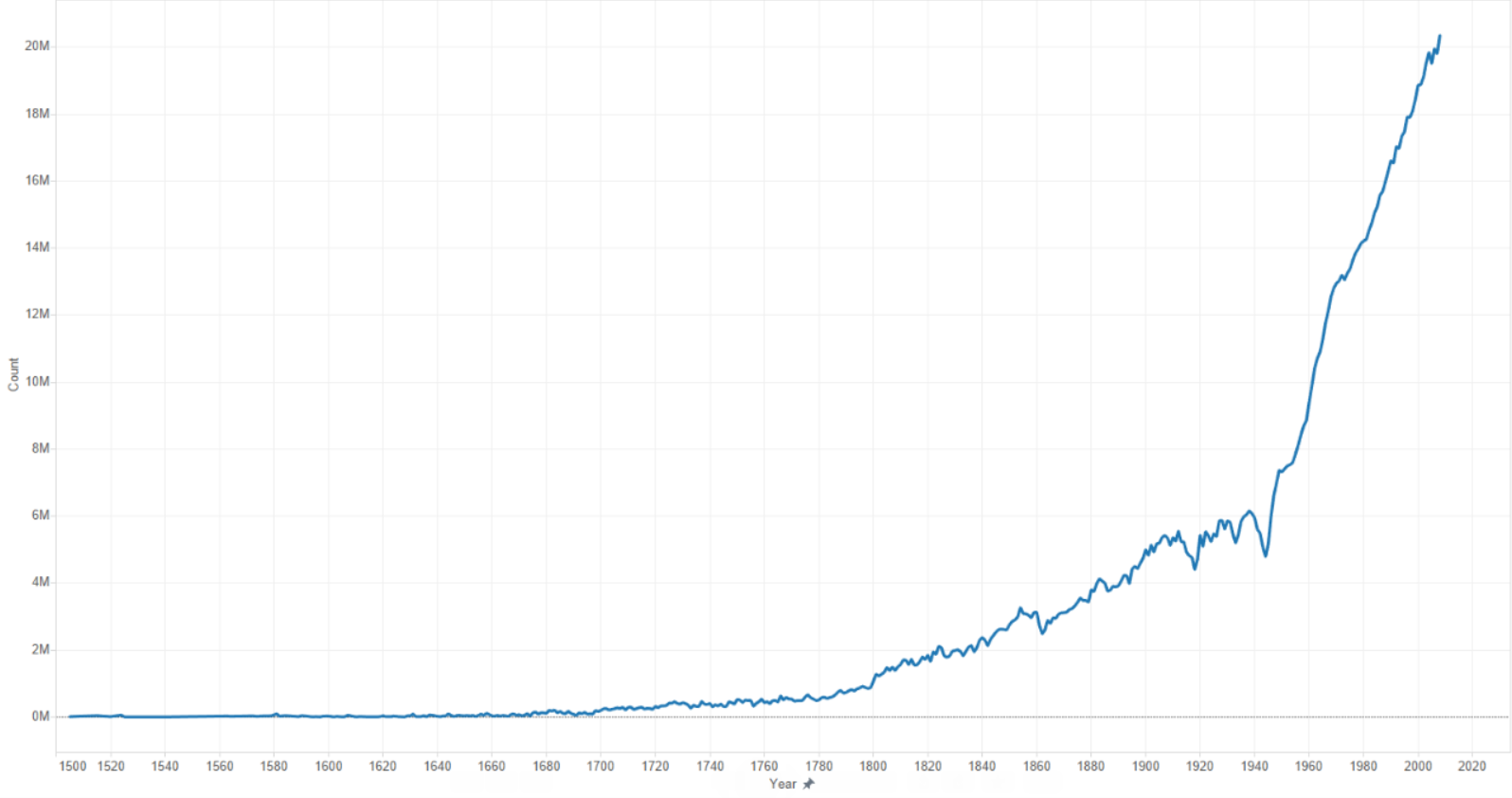
I have used Hadoop streaming to get to the solution

* In the Mapper program, each line of text is read and words in line are converted to lower case and split by white space by removing punctuations.
* Year is taken as key and word is taken as value.
* Mapper function will issue (year,1) as a key-value pair, ex: (1505,1)
* All key-value pairs will be sorted and will serve as input to reducer.
* For each line, reducer will split key and value by tab and for each key, all values are summed up to display final key-value pairs as below

Example: 1505 11055

1506 18569

**Year – Word Graph:**



**Twitter Questions 1 and 2:**

**Required data to answer the questions:**

1. Number of days in data.
2. Number of days on which PrezOno twitted.
3. Number of tweets per each hour.
4. Number of different week days (example: number of Mondays in data, number of Tuesdays in data etc.)
5. Number of different week days on which PrezOno Tweeted. (Example: Number of Wednesdays on which he tweeted.)
6. Number of tweets on each week day.

From the above requirements it is clear that if we calculate Number of different week days and sum them all we get number of days. Therefore it is enough to calculate 4 to get 1 in above requirements.

Moreover, if we calculate number of different week days on which PrezOno tweeted and sum all of them we get number of days on which PrezOno tweeted. Calculating 5 gives 2.

By solving two questions together we can verify results comparing sum of 3 to sum of 6 (Total tweets by PrezOno.)

**Approach:**

We used Hadoop streaming method to solve these questions. The steps followed in mapper are as below:

1. Mapper parses each input line using json library and extracts the ScreenName, Created\_at. Created\_at is used to calculate week day (example: Wed), date (example: 24Sep2014), hour (example: 21).
2. The dates are added to a set. This set is a value in dictionary where key is week day. Since the set contains only unique values we get the unique values for each week day.

**Example**: {'mon': set(['22Sep2014', '29Sep2014']), 'wed': set(['24Sep2014', '01Oct2014'])}

1. If the user is PrezOno we do the same thing as step two to another map. This map contains the week days on which PrezOno tweeted. The key value pair (weekday+’Tweets’,1) and (hour+’Tweets’,1). Sum of these values for each key from all the mappers gives us 3 and 6 in requirements.
2. After all the lines in input to a mapper are exhausted we print all the dictitems in maps in step 2 and 3 as key ‘x’+weekday and value as the corresponding set. We prepend with ‘x’ to differentiate the key value pair containing sets to decimals.
3. Reducer add all the values corresponding to a key if the values are decimals and output the key value pairs or performs an union operation on sets when the values are sets and outputs key as key and value as size of the set.
4. We run this with single reducer and no combiner is used as we are doing much work in mapper.

**Results: Question1**

**Total Number of days in data**: 313

**Total Number of days on which PrezOno twitted**: 192

**Tweets per each hour**:

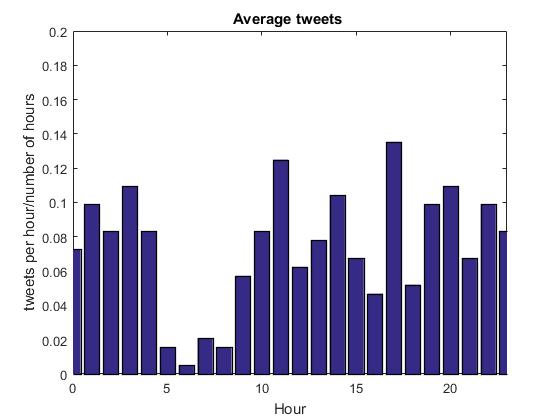
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **00** | **01** | **02** | **03** | **04** | **05** | **06** | **07** | **08** | **09** |
| 14 | 19 | 16 | 21 | 16 | 3 | 1 | 4 | 3 | 11 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** |
| 16 | 24 | 12 | 15 | 20 | 13 | 9 | 26 | 10 | 19 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **20** | **21** | **22** | **23** |  |  |  |  |  |  |
| 21 | 13 | 19 | 16 |  |  |  |  |  |  |

**So he tweets most on average at 1700-1800 or 5pm to 6pm**

Plot of the expected number of tweets for each hour of the day, for those he did tweet is as shown below: X-axis has hours and Y-axis has Average (Number of tweets in hour/Number of hours).



**Results: Question2**

**Number of week days on which PrezOno tweeted:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 28 | 20 | 31 | 28 | 23 | 31 | 31 |

**Total number of week days:**

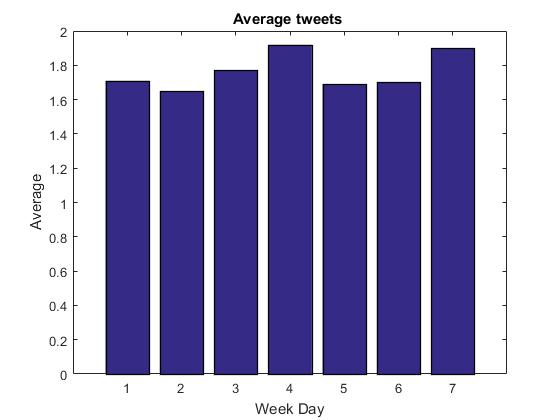
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 46 | 46 | 43 | 46 | 42 | 44 | 46 |

**Number of tweets per week day:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 48 | 33 | 55 | 54 | 39 | 53 | 59 |

**Average Tweets per week day:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 1.71 | 1.65 | 1.77 | 1.92 | 1.69 | 1.70 | 1.90 |

****

**Twitter Question 7:**

**Required data to answer the questions:**

1. Average of Left Co-ordinate of geo component
2. Average of Right Co-ordinate of geo component
3. Number of tweets having geo data
4. Number of tweets having no geo data

**Approach:**

We used Hadoop streaming method to solve these questions. The steps followed in mapper are as below:

1. Mapper parses each input line using json library and extracts the geo data if present.
2. If geo data is present, then ‘yes’ value is incremented and left and right coordinates are added to the present coordinates.
3. If geo data is not present then ‘no’ value is incremented.
4. After all the lines in input to a mapper are exhausted we print all the variables as key and value pair.
5. Reducer add all the values corresponding to a key.

**Results:**

left 72845727.274578

no 4212000.000000

right -154046999.285180

yes 1868302.000000

**Centroid = ((left/yes),(right/yes))**

****

**Proportion of Tweets:**

Proportion = yes/no = 1868302.000000/4212000.000000 = 0.4435

***Twitter data question 3:***

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

***Required data to answer the questions:***

1. User data in twitter data set
   1. Only screen name is required to answer the question.
2. Twitter text – Only length of the text.

According to the question, we get to know that we have to find tweet length of each person as well as President Ono’s tweet length. At the end we have to calculate the average tweet length of President Ono and a cumulative tweet length of all others. This can be done using map reduce by following the below approach.

***Approach:***

Using Hadoop streaming method we can solve this question. Find below the steps followed in solving the question:

1. Initially, we create a mapper which parses each input line using json library and extracts the Screen Name from user data and Twitter text.
2. In the mapper we iterate through each user along with his tweet from the beginning and when the mapper encounters the screen name of President Ono, it calculates the text length as well as it increases the counter of each tweet.
3. When the mapper encounters individuals other that President Ono, it calculates the text length as well as it increases the counter of each tweet.
4. Mapper sends the output to the reducer and reducer considers them as a key value pair for President Ono’s tweets and other’s tweets.
5. Reducer sums up the all the values corresponding to the specific key and performs the union operation on them. At end we get total tweet length text of president Ono along with the number of tweets and also we get the total tweet length of all others (cumulative data) along with the number of tweets.
6. Used only one reducer and a combiner to complete the job faster.
7. We can calculate the average twitter text length of president Ono when compared with all others (cumulative data of others)

***Results:***

ocount 6079961

osum 496716166

pcount 341

psum 35578

***Final result:***

Average twitter text length of President Ono – psum/pcount = 35578/341 = **104.33**

Average twitter text length of all others (cumulative data) – osum/ocount = 496716166/6079961 = **81.697**

*When compared to other’s president Ono’s tweet text length is* ***more than*** *all other’s tweet length*