CLOUD COMPUTING P3

Question No. 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. - Suraj Kamble

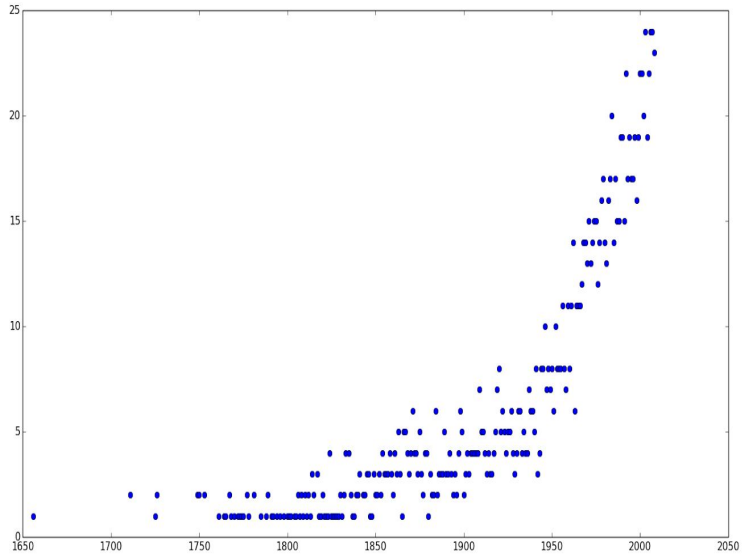
- We are asked to count the total number of words found in google 1-gram corpus, without repetitions, for each year available. The 1-gram data consists of 4 tab separated fields per row; the first field is the word, second is the year in which the word occurred, third field is the total number of times the word was found in the year(second field), and the fourth field is the number of documents the word was found in.

For this particular problem, we need the length of the set(a set is a list of items without repetitions) of words found in each year. We can achieve this by just using the second field from 1-gram data. If we assign count '1' to each occurrence of a year in the data, we can eliminate repetition of the words. For example, if the data is:

Linus	1995	4	2	
Torvalds		1998	6	3
Jobs	2000	5	2	
Steve	1995	4	3	
Gates	1999	8	4	
Bill	2000	9	4	

Here, to find the length of the set of words found in each year, we just count the number of occurrences of a year. So, for 1995 the length is 2, for 1998 it is 1, for 1999 it is 1, for 2000 it is 2.

I have used Hadoop-Streaming mode for this problem. The mapUWC.py program assigns count '1' to each occurrence of an year and the reduceUWC.py program sums up the count for each year and plots the graph accordingly.



RESULTS: Following is a part of the output, when the program was run on 1-gram data.

...
..

1997	19
1998	16
1999	19
2000	22
2001	22
2002	20
2006	24
2007	24
2008	23

...

Question No.2. How does @PrezOno's tweet length compare to the average of all others? What is his average length? All others? - Sarat Chandra Lingamarla.

We are asked to compare the average tweet length of President Santa Ono versus average tweet length of rest of the twitter users. The users object has a screen_name field which is used to filter president Ono's tweets from the twitter data.

The number of tweets and the sum of tweet lengths is calculated for President Ono as well as all other users by iterating through the data.

I have used Hadoop Streaming mode for this problem. I ran the program on a part of the twitter data before running it over the entire data. The maptw.py program iterates through the data, compares the screen_name field of JSON object with 'PrezOno' and increments tweet length and tweet count .Similarly, it also calculates total tweet length and count for other users. The reducetw.py program takes mapper outputs and aggregates the tweet lengths and tweet count. It then calculates the sum and average tweet length of PrezOno and others.

Result:

PrezOno's average tweet length:104

Average tweet length of all other users:83

The results are rounded to the nearest integer.