Problem:

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

Result:

President Ono’s average tweet length: 104

Average tweet length of all others: 81

Analysis:

This problem is solved using spark. The data is read using streaming mode. The ‘getText’ function is used to extract the tweet details such as tweet length. The ‘getText’ function takes parsed JSON as input and each JSON object is validated with username fields to find President Ono’s data and tweet length is extracted. Similarly, average tweet length of all other users is calculated.

Using reducedByKey created rdd’s of tweet length and tweet count. Average tweet length of PrezOno and remaining others by using the calculated tweet count and tweet length values and printed results. Results can also be saved to a file using saveAsTextFile function.

CLOUD COMPUTING P4

Question No. 1. Plot the average word length for all unique words for all years available. Year on
xaxis, average wordlength on yaxis. - Suraj Kamble

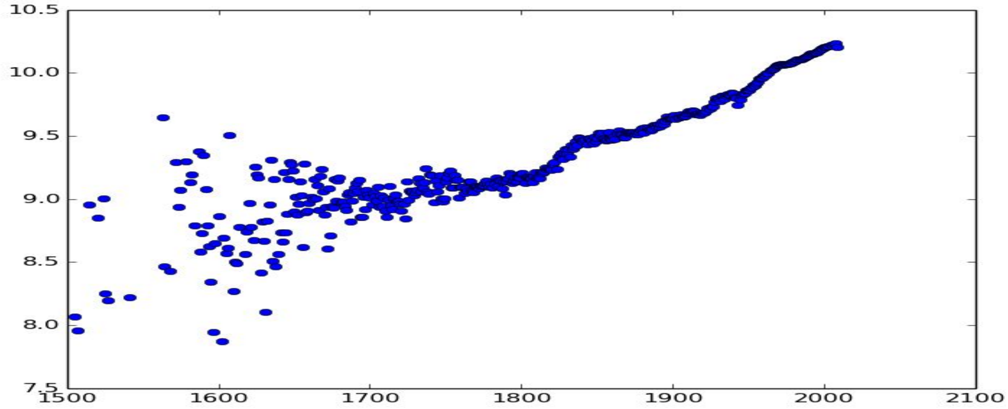
- First we count the total number of words found in google 1-gram corpus, without repetitions, for each year available, then we add the lengths of all these unique words for each year.

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 the map() function, we provide the year as key and a list containing the length of word and “1” as the value to the key. 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. For key 1995, the value is (5, 1) for each occurrence. We aggregate all these occurrences using a reduceByKey() function to add all the lengths and the number of occurrences. We use another reduce function to find the average of the word lengths per year.
<u>Results</u>:


10.5
10.0
9.5
9.0
8.5
8.0
7.5
1500
1600
1700
1800
1900
2000
2100