**Project P3 Report**

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

1. Plot the average word length for all unique words for all years available. Year on x-axis, average word-length on y-axis.

**Ans:**

This question requires average word length for all unique words for all years available.

**Approach:**

I have used Apache Spark to get to the solution

In input file, each line has data with word, year it has occurred in, number of times it occurred

in a year.

Each line of input is split by tab.

Mapper emits (year, (wordLength,1)) key value pair

Now we will send year as key and wordLength as value from mapper

Reducer will sum up all wordLengths for each year.

Now mapper will emit (year, (wordLength/total number of words)) key value pair

Finally, we save the result in a text file

**Graph:**

