**Google 1gram question 1** (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.

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

**Approach:**

1.       Map all the words in the data in the format (year, 1)

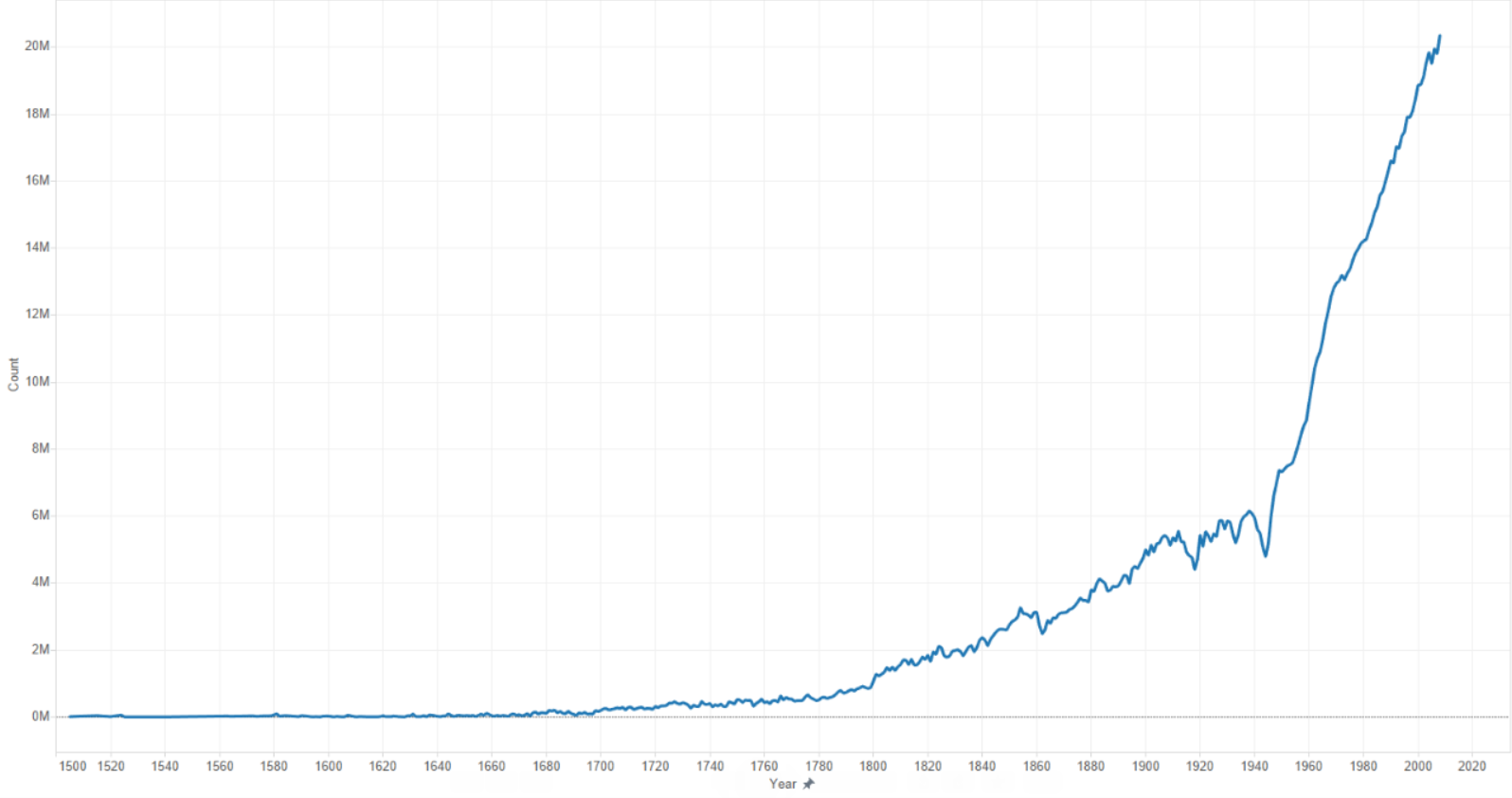
2.      Apply Reduce operation using year as key to get the number of words in year.

3. Plot the graph with year on x-axis and number of words on y-axis

Example: 1505 11055

1506 18569

**Year – Word Graph:**



**Google 1gram question 3**

3. 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

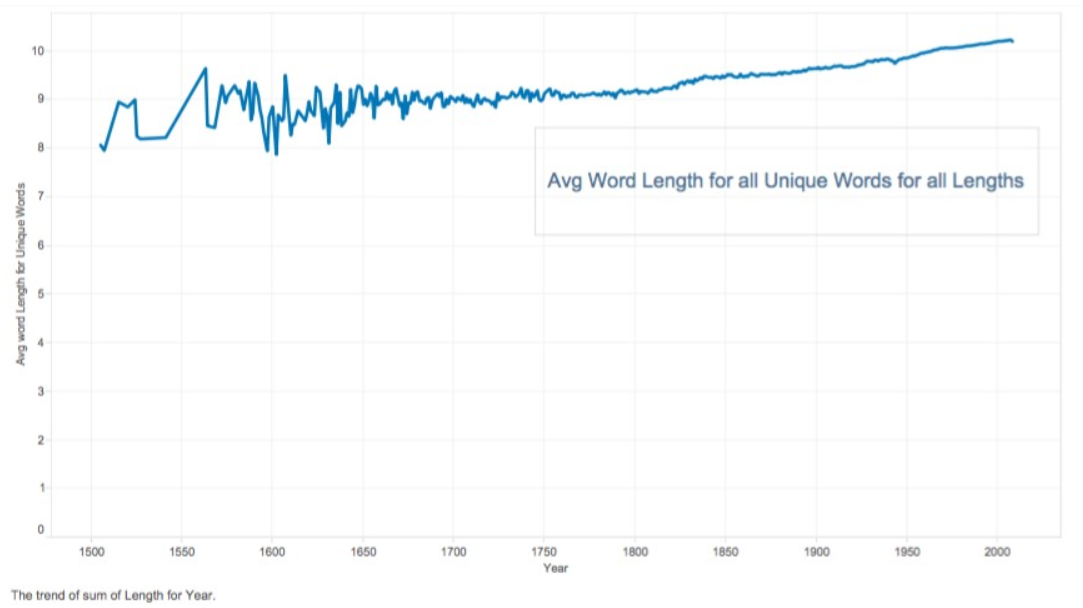
Now we will send year as key and wordLength as value from mapper

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:**

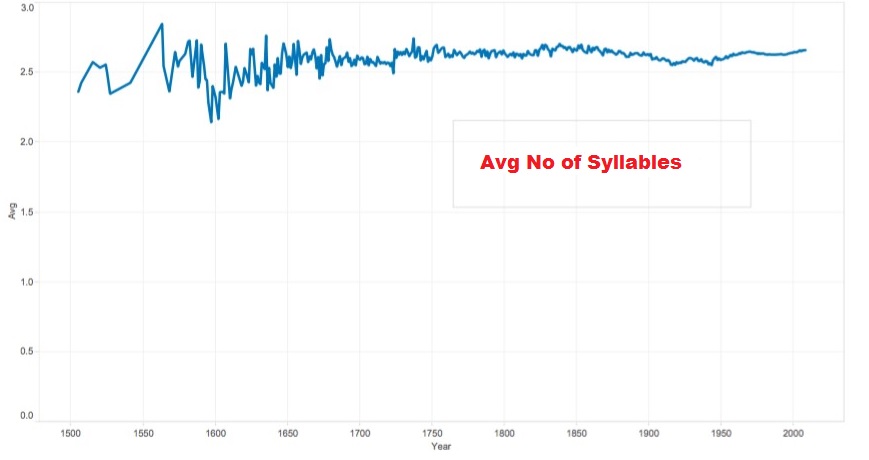


**Google 1gram question 4:**

4. Plot the average number of syllables on the y-axis, and time (year) on the x-axis.

1. Map all the words in the data in the format (year, syllables per word).
2. Reduce by key, year in this case to get the sum of all the syllables in a year.
3. Use the results from problem 1 in 1gram data to get number of words in a year.
4. Get the average and plot the graph.

Results:



***Twitter data question 5:***

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

***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. At the end we have to write 3 different queries to calculate the user who tweeted the most, top 5 longest tweeters and bottom 5 tweeters over each’s average tweet length. This can be done using Spark by creating a python file (twitter 5) and submitting it to Spark by following the below approach.

***Approach:***

Using Spark and writing a simple python code we can solve this question. Find below the steps followed in solving the question:

* Initially, print\_function, SparkContext, SQLContext are the main functions that have to be imported to the python file.
* Then a function is created a mapper which parses each input line using json library and extracts the Screen Name from user data and Twitter text.
* The dataset is then loaded into a RDD using the sc.textFile(sys.argv[1]) when the file name is sent from the command line parameters.
* Using “flatMap” the RDD is then transformed by using the function getInfo which was initially written to parse each input line.
* Using “sqlContext” and “createDataFrame”, a map is implemented to create a new transformed RDD as shown below

sqlContext.createDataFrame(texts.map(lambda (u,t): Row(text = t, length = len(t), user = u)))

* A temporary table is created using the function registerTempTable(“Name\_Of\_Table”).
* 3 SQL queries are written to find out the user who tweeted the most and the top/bottom users from the temporary table created in the above step.
* The program is then sent to the Hadoop cluster using the following command:

spark-submit --master yarn-client twitter5.py hdfs://hadoop2-0-0/data/twitter/part-03212

* At the end the results are printed using few for statements and then printed in the console.

***Results:***

marilyn9743 3419

Huntersweat 416

RoyalEliteKiva 350

blackxhole 320

KelleeMichele 272

pizzadellarry 253

Laila\_Lafrai 0

Fun\_Size20 0

Trevorsturgill5 0

2013Afi9 0

lm\_Lil\_Wanie 0

***Final result and observations:***

The first line is the output for twitter user who tweeted the most who is **marilyn9743** and the tweet length = **3419 characters**

The second set of output (next 5 lines) are the top 5 longest tweeters over each’s average tweet length. We can observe that the 1st longest tweeter is approximately double the 5th longest tweeter.

The third set of output (next 5 lines) are the bottom 5 tweeters over each’s average tweet length, we can observe that their tweet length is empty which means they just submitted a tweet with length of the tweet as 0.