SoftDes Text Mining

Byron Wasti

February 2015

1 Overview

For this project I decided to focus on learning how to manipulate large amounts of data, and access the large amounts of data repeatedly without doing a lot of processing. What this meant is I would have to focus on making a database and a type of cache system in order to store data in case of repeated access.

I decided to compare the average length of words in a Wikipedia entry for various topics, and see if there were any trends.

2 Implementation

There were three main components to my project. The first, and perhaps most simple, is the database I used in order to eliminate the need to recalculate the average length for topics every time I ran the program.

I decided to use the anydbm module, which easily stores a dictionary object in a text file, and can load it just as easily. When I called my Averaging function, if the word is found in the dictionary (meaning it had already found the average length of words) it would return the value immediately, saving time. If the word was not found, once the average length of words for that topic is found, it is stored in the database.

The second component is the Averaging function. This used pattern's Wikipedia accessing abilities, which are phenomenal. The function goes through each section of the Wikipedia article, splits up the selection into individual words with .split(' ') and then averages their lengths and returns this average.

The final component is graphing the data, such that trends could be more easily seen. I used the MatPlotLib module in order to do graphing, and made bar plots with the x-axis being the labels of the different topics.

3 Results

The topics I ended up searching I grouped into different topic groups. The main groups I had were animals, colleges, countries, states, sports, academic topics, long words and short words. I wanted to check the average length of words for the top 5 longest words and a lot of short words because a few people questioned the validity of my scientific work as things that are inherently a long word could skew the results.

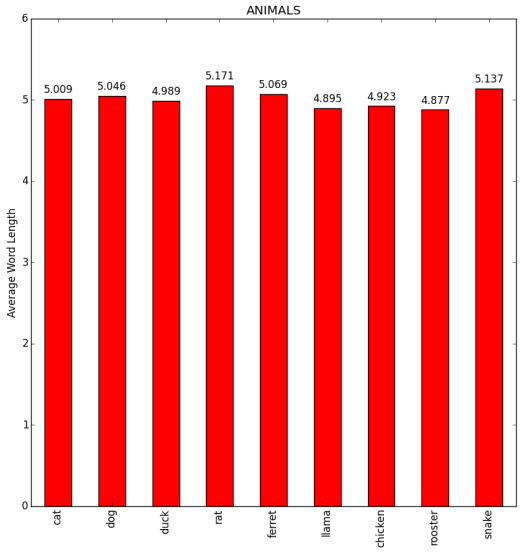

Figure 1: Average length of words in different animal's Wikipedia entries. They seem to average around a length of 5 characters.

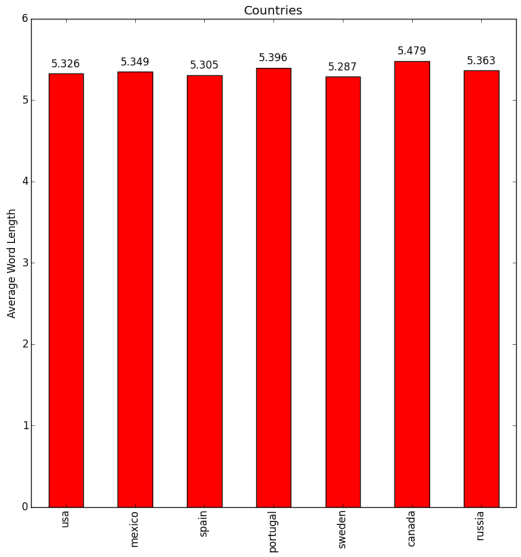

Figure 2: Average length of words in different countries's Wikipedia entries. They seem to average around a length of 5.3 characters, which is longer than animals but not significantly so.

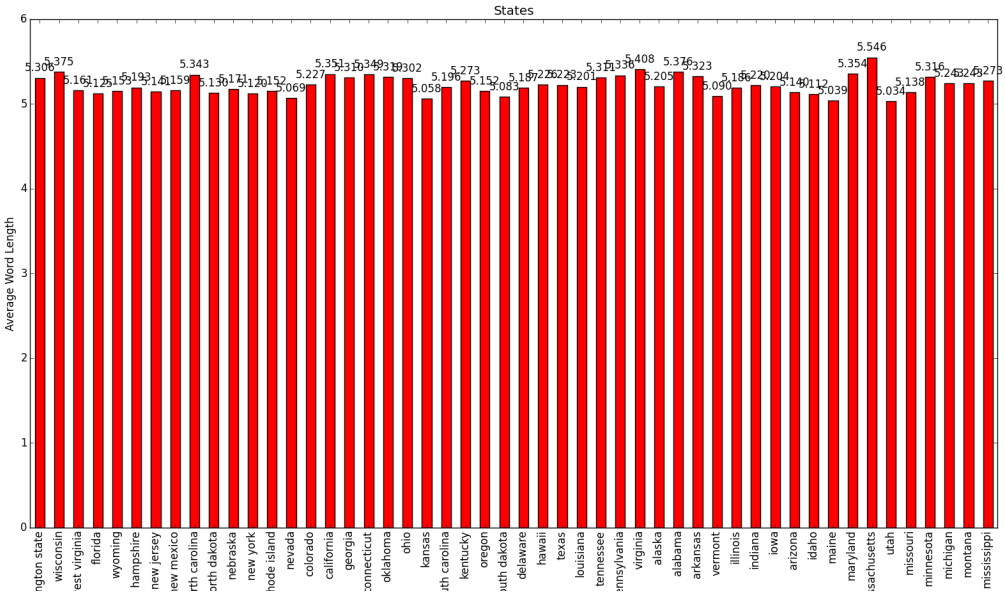

Figure 3: Average length of words of all the state's Wikipedia entries. This one is fun because it varies quite a bit. It would have been nice to see if the southern states have different average lengths than northern states, but organizing the states was out of the scope (ie. time consuming) of this project.

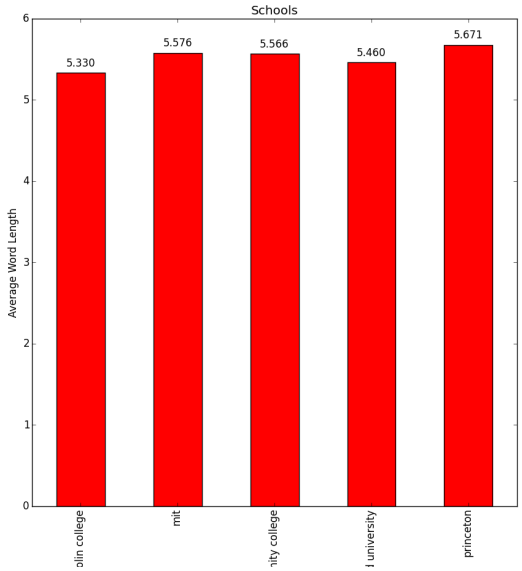

Figure 4: Average length of words in different school's Wikipedia entries. Sadly Olin is not in the lead...

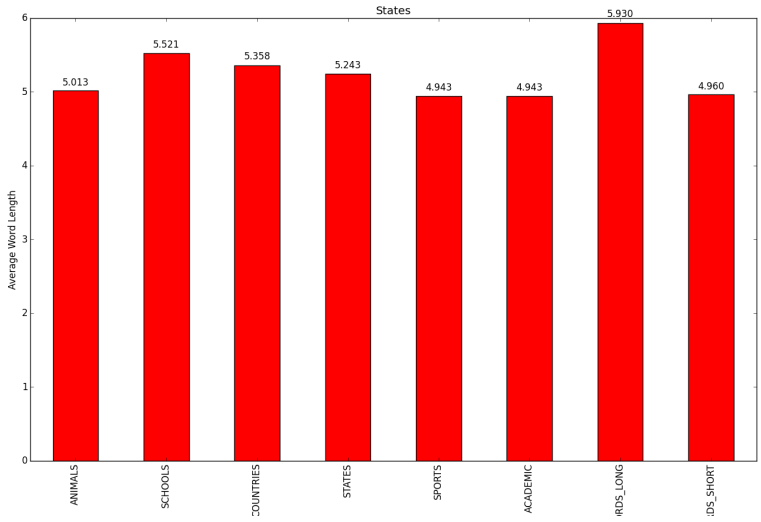

Figure 5: Average length of words in different general topic area's Wikipedia entries. The top 5 or so longest words really take the lead, with everything else being around the same, aside from colleges which is a little higher.

From the graphs it is not very apparent that there are any trends. On average it seems that most words are around 5 characters long, and this seems to be rather pervasive in all the Wikipedia entries. One thing to note is that College wikipedia entries tend to have longer words used in them then those of animals or sports. This does not necessarily mean anything, but it is interesting to think about how different Wikipedia entries may include longer or shorter words depending on the audience.

It does also seem to be the case that the average word length in Wikipedia entries for absurdly long words is higher than average, but it is not that different from normal word lengths. Therefore I think it is safe to say that the length of the topics word does not actually impact the average word length in the Wikipedia's entry all that much.

4 Reflection

While my results sadly don't show any prevalent trends in the average word length of Wikipedia's entries, it did teach me that the average word length in at least Wikipedia's entries are around five characters. Although my original plan was to translate Wikipedia entries via Google Translate and then compare them to native Wikipedia translations, that fell through because you have to pay for Google Translate's API.

After spending a while trying to bypass Google using a variety of methods, I decided to just come up with a slightly lamer idea. I am, however, happy with my implementation, and I learned how to use the anydbm module which is super awesome and will be used later.