Milestone Report

Mikhil Raj

29-03-2015

Abstract

Portable office actually means the works done on the cellphone and the tablet and we need input system to saving our time on typing on them. So a smart and efficient keyboard is required and the core of this input system is a predictive text model. The objective of this report is to analyze the data for the Capstone Project. Therefore I analyzed the 3 data sets "twitter", "blogs" and "news". I will show that each data set have a different usage of language.

As this report is supposed to be read by non-data-scientists you will not find any R-Code in it. If you are interested in the R-Code you can see it in my (github account).

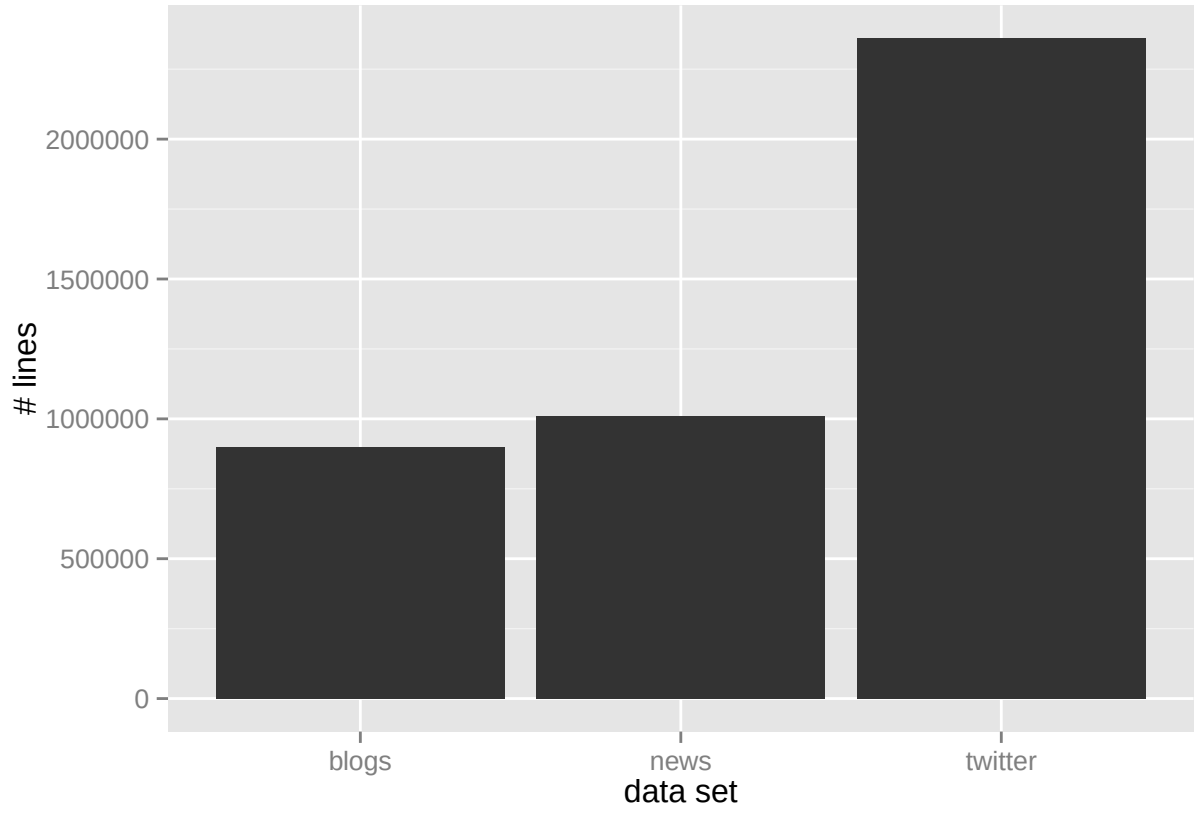
Data Collection

The data were downloaded from the course website (from HC Corpora) and unzipped to extract the English database as a corpus. Three text documents from the twitter, blog and news were found with each line standing for a message.

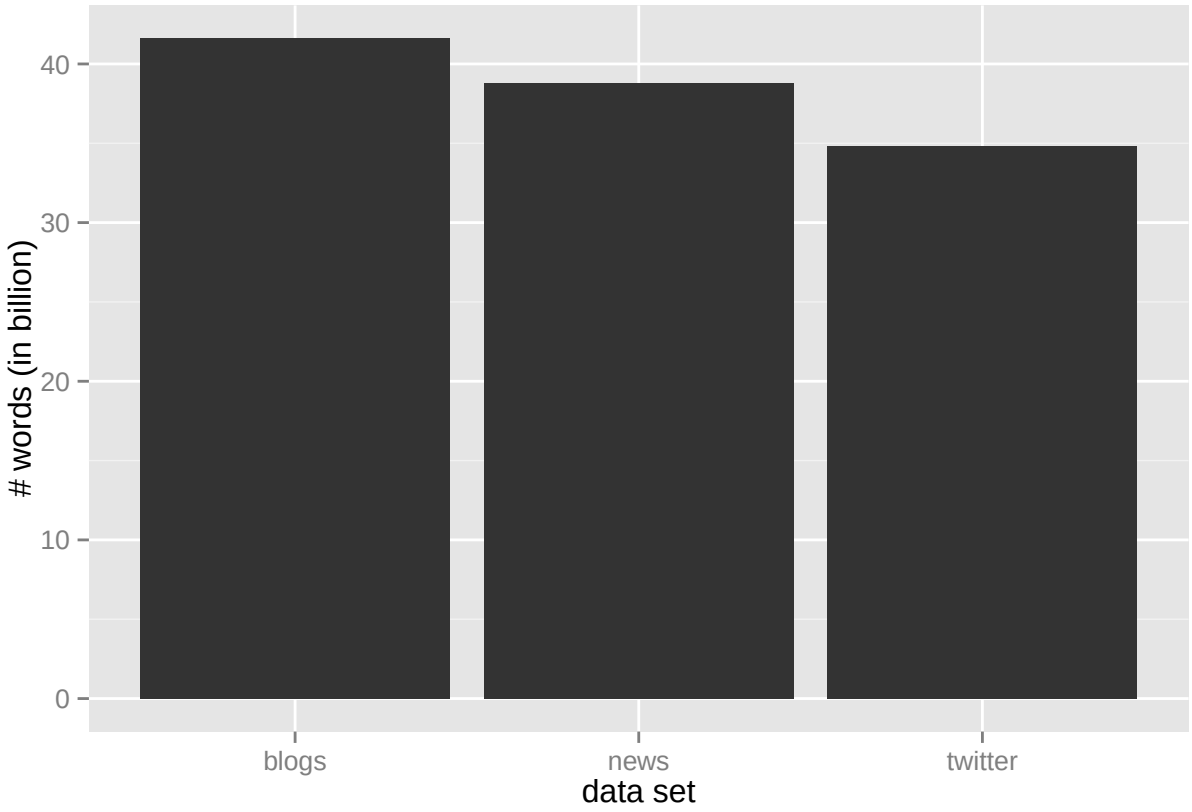
Exploratory Analysis

First of all, I cleaned the data by tokenizing the words by any punctuation / whitespace.

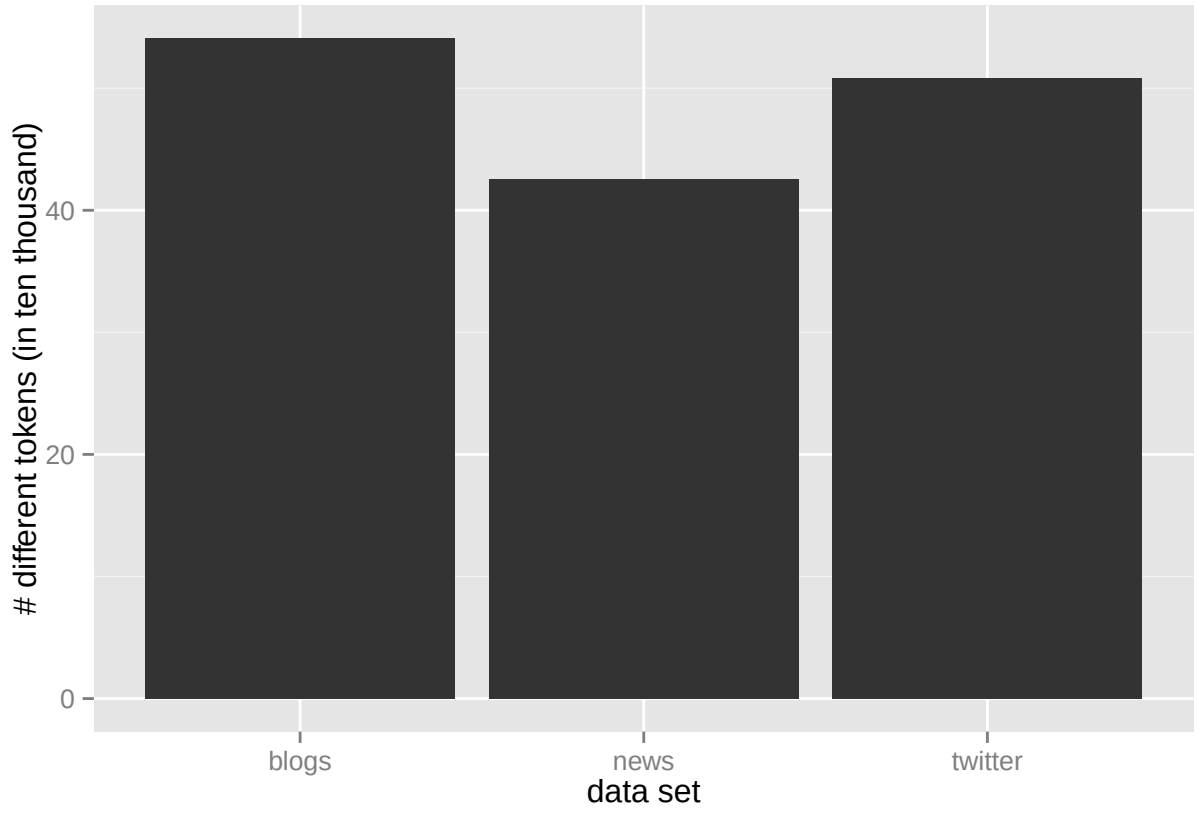
First, let's have a look at the number of lines of each data set. One can see that with 2.3 billion lines the "twitter"-dataset has the most number of lines. The second most is news with about 1 billion and blogs has 900,000 lines.



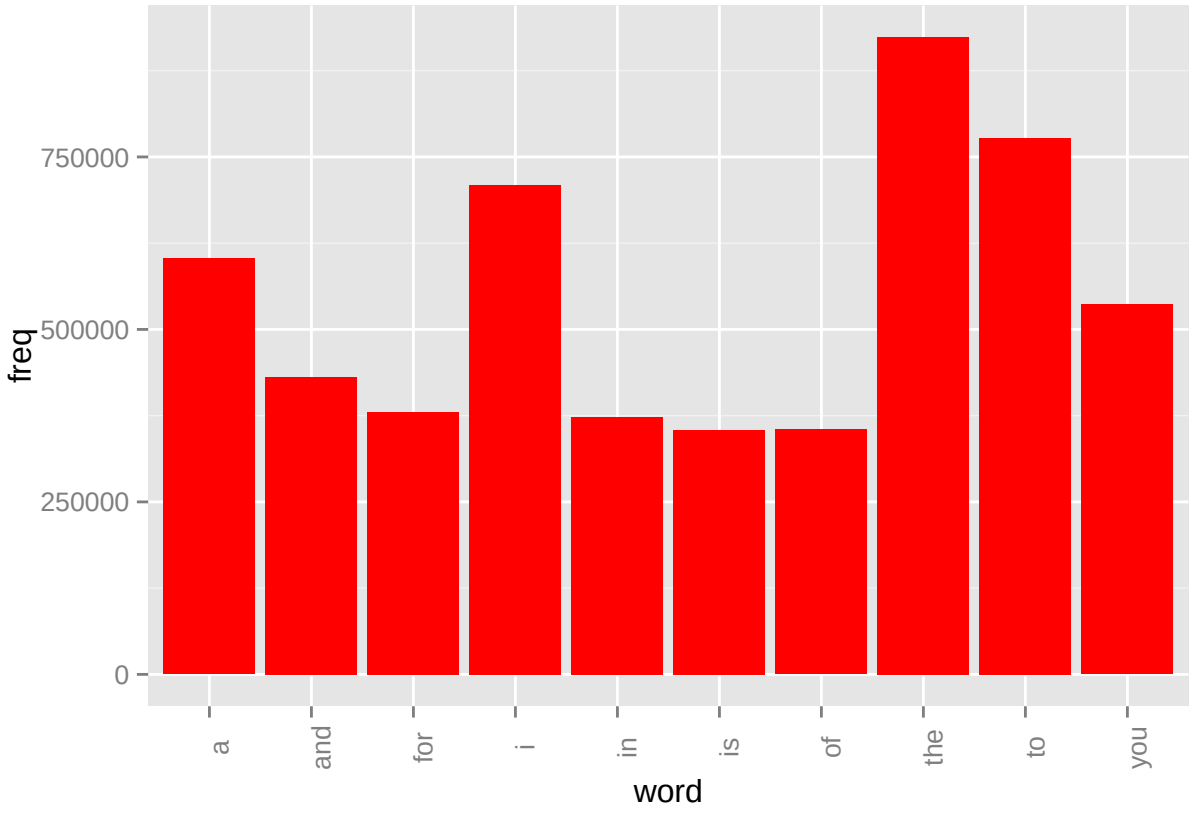
One might suggest that the one dataset with the most lines of code also has the most number of words. But as can be seen below, the twitter data set has the lowest number of words. This makes sense as the maximum number of characters per each tweet is 140. Both news and blogs have higher number of words with around 40 billion.



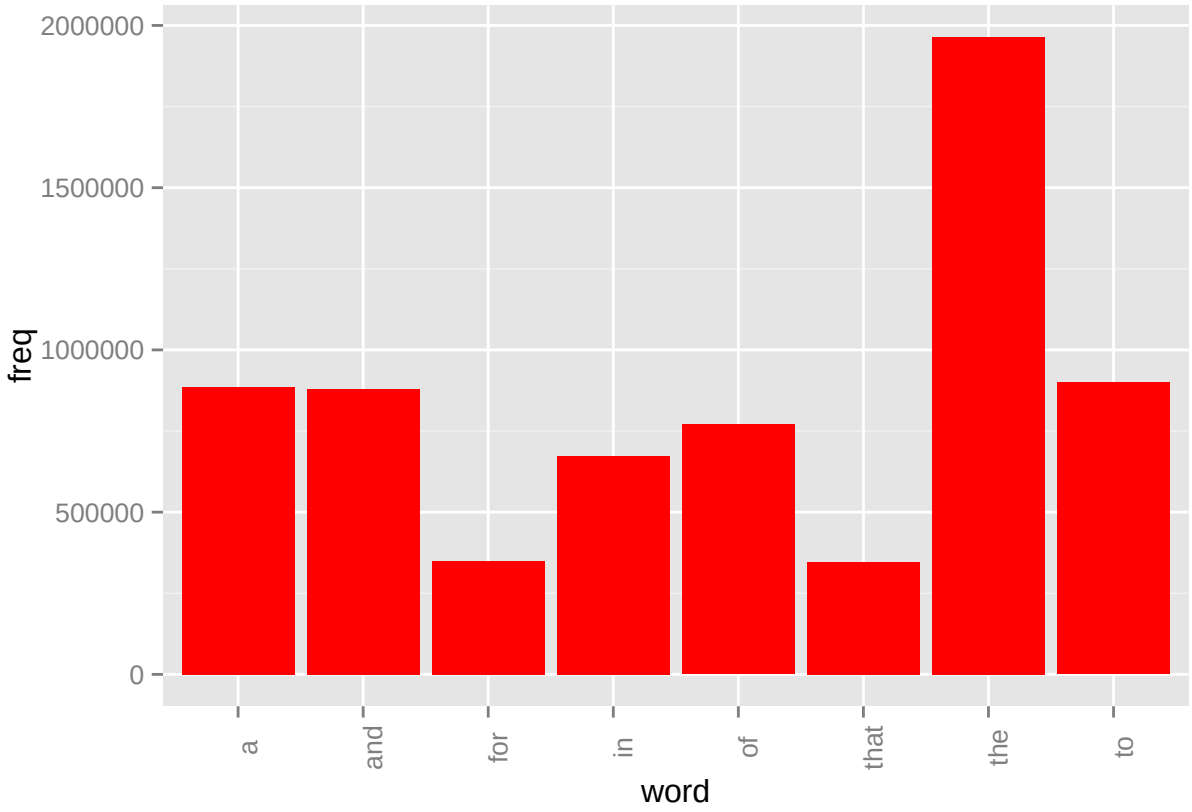
Surprisingly the number of different tokens is the same for blogs and twitter. I was supposing that people who are using twitter would use less different words than people that write blogs. Also, I was supposing that professionals who write articles for news would have the most vocabulary.



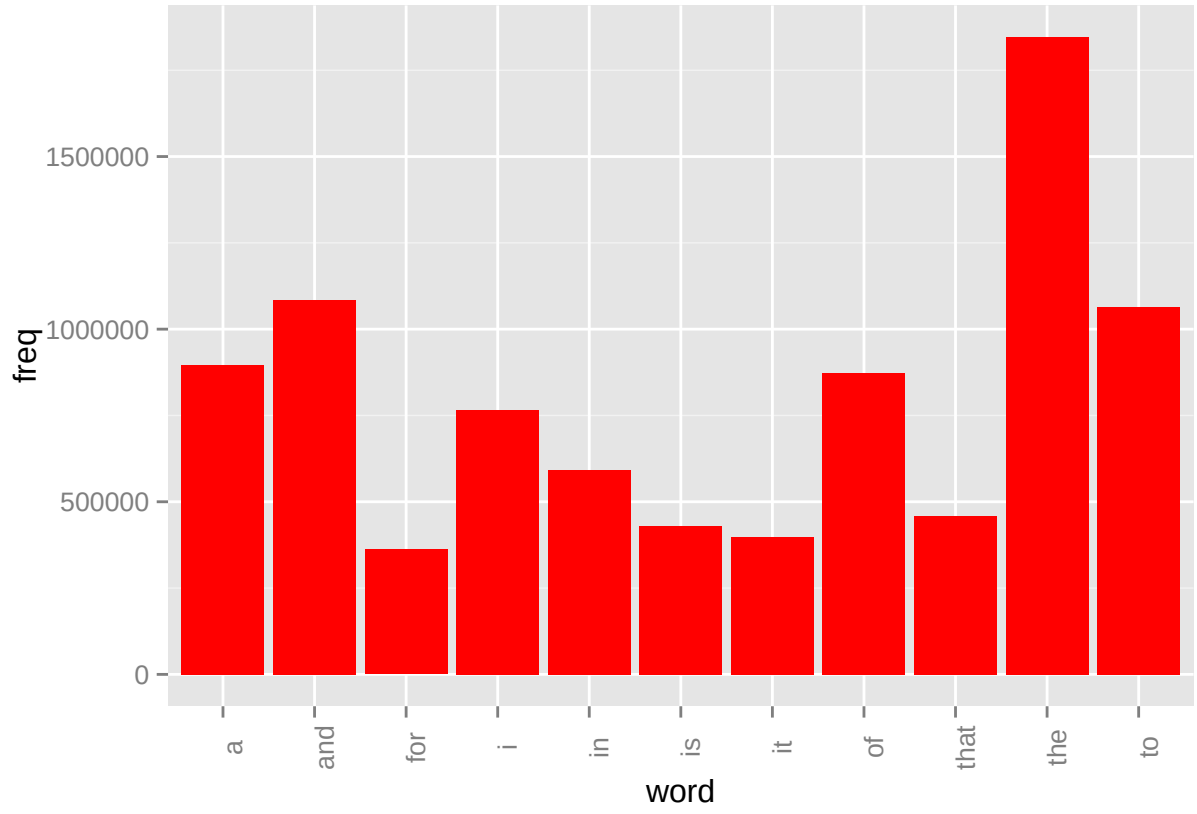
When you have a look at the most frequent words of the twitter dataset besides the words that are used for conjunction like “and”, “or” and “to” the words “I” and “you” have a high frequency: These are the most frequent words in the twitter dataset:



These two words naturally aren't that frequent in the news data set:

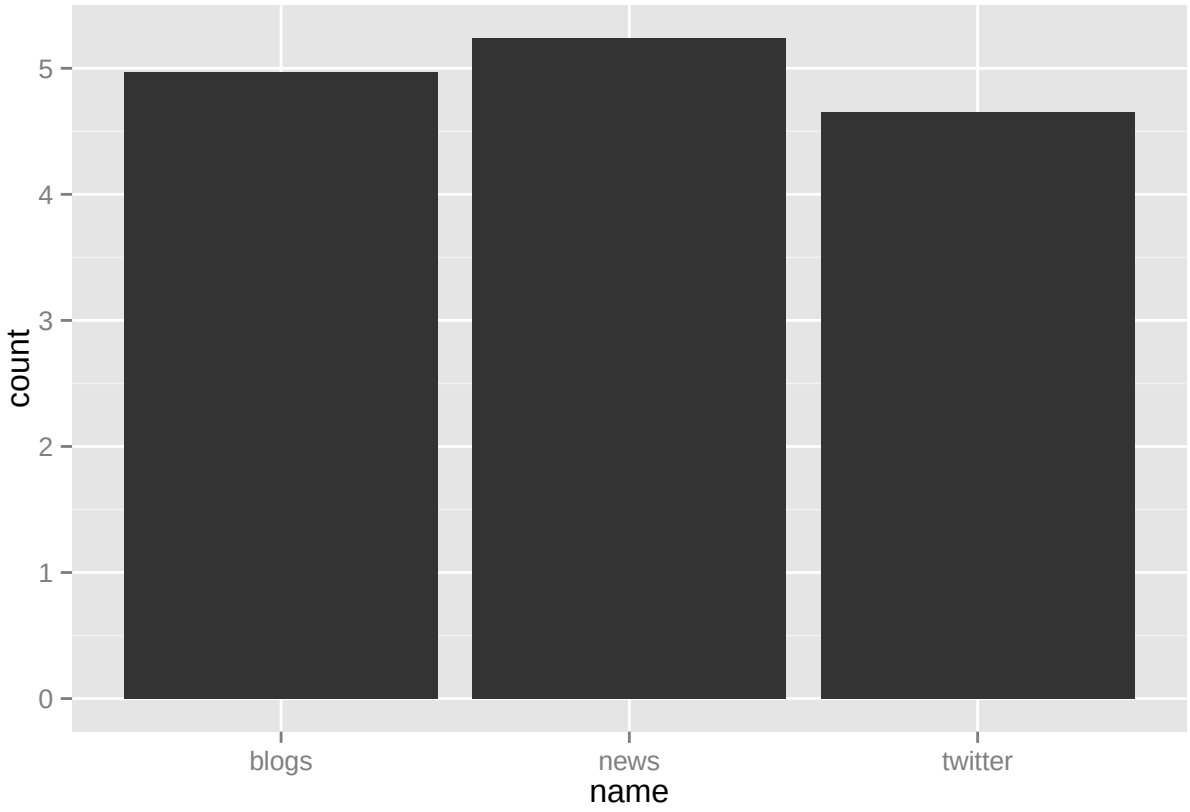


The same is for blogs:



What is also interesting is the average number of characters per word. Due to its restriction, people that are using twitter are using shorter words (4.59 characters per word) and news have an average of 5.23 characters per word.

```
## [1] 4.651 4.971 5.237
```



Conclusion

I will probably learn separate models for each dataset, since people use language differently in them. Until now my code works good and is able to process all data sets complete. I will investigate how many number of lines I would need for training and test set and see if I should use another NLP package.

Future work

- Evaluation of other NLP packages or parallelization of own code.
- Create test and train datasets and set error rates and confidence level
- Clean data of profanity words
- Decide on n-Grams
- Evaluate the limitations of shiny app