DS-Capstone-Milestone-v1

March 20, 2016

Load required R Libraries for Natural Language Processing and set working directory and file path.

Load data This routine uses data from a corpus called HC Corpora (www.corpora.heliohost.org). The readme file at
http://www.corpora.heliohost.org/aboutcorpus.html provides details on the corpora available. The files have been language filtered but may still contain some foreign text. This routine applies data science to natural language processing.

The first calculation reads a large twitter data file, a large news file, and a large blog collection.

```
##                      Lines
## "lines of twitter"  "2360148"
```

```
##                  Lines
## "lines of news"  "77259"
```

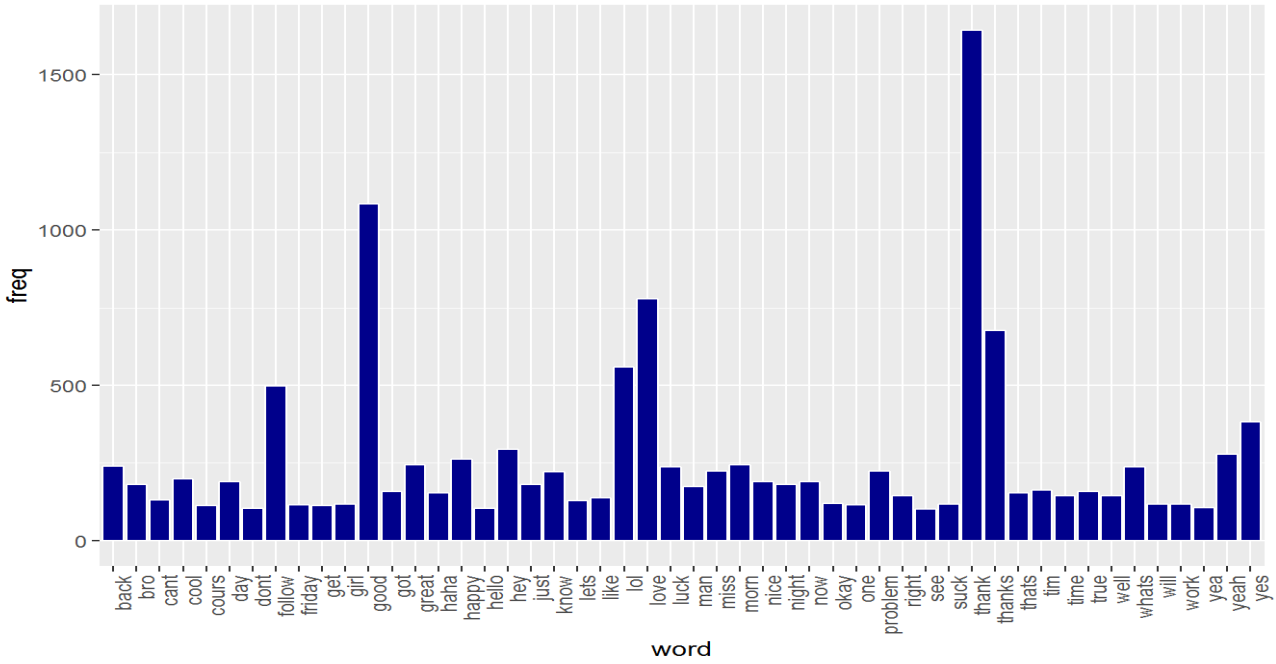
```
##                      Lines
## "lines of blog"  "899288"
```

The next calculation provides statistical calculations on the data characteristics.

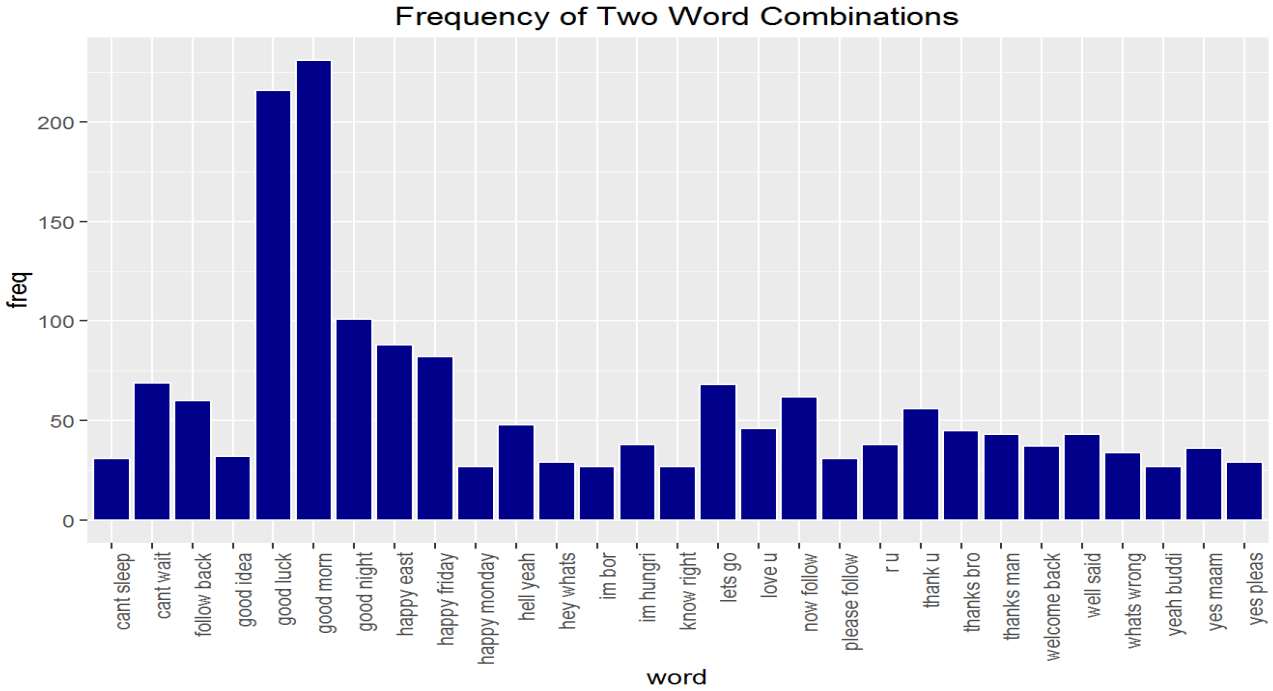
```
## <<DocumentTermMatrix (documents: 3, terms: 5961)>>
## Non-/sparse entries: 6531/11352
## Sparsity           : 63%
## Maximal term length: 16
## Weighting          : term frequency (tf)
```

The most frequent words in the combined dataset are depicted in a word cloud.

The chart below illustrates the frequency of single words in the combined dataset.



The chart below illustrates the frequency of paired words in the combined dataset.



The chart below illustrates the frequency of triple words in the combined dataset.