Milestone for Capstone project

Milestone for Capstone project

The goal of this report is to perform an exploratory analysis, explain the major features and outline my plans for the eventual app and algorithm.

It contains basic summaries such as word counts, line counts and others as well as different plots to better represent the data.

I used three data sets for this analysis containing: blogs, news and twitter data.

R code chunk publication has been avoid to allow that a a non-data scientist person understand it.

Basic summary

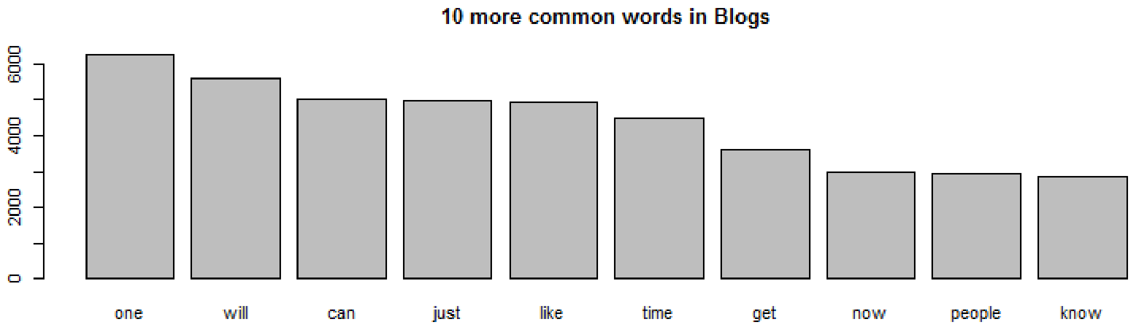
The following table shows the number of lines and words in each file as well as the average number of characters in the entries of each file:

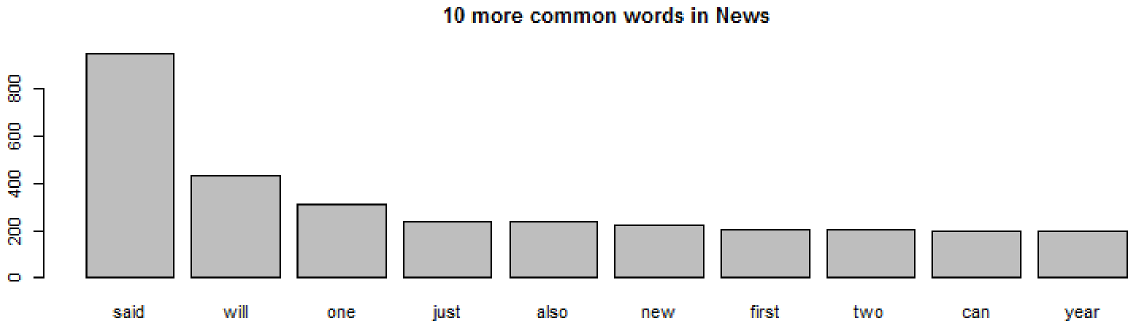
```
Dataset             Number of Lines   Number of Words     Average number of characters
en_US.blogs.txt     899,288           37,541,795          231.7
en_US.news.txt       77,259            34,762,303          203
en_US.twitter.txt   2,360,148         30,092,866          68.8
```

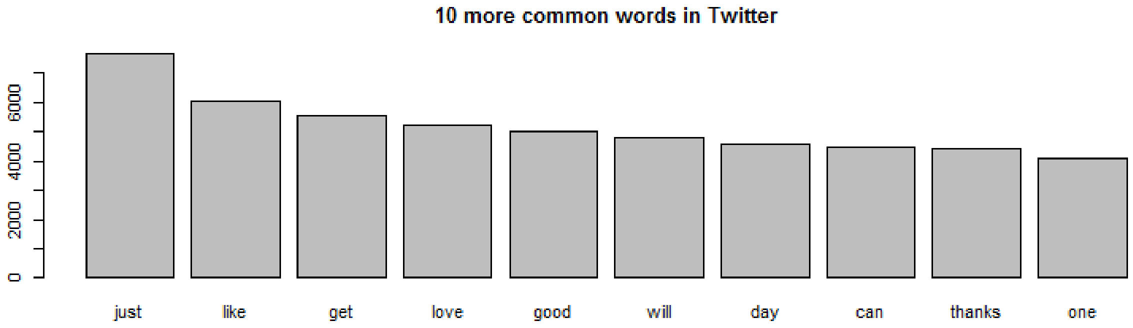
Data cleaning and exploratory analysis

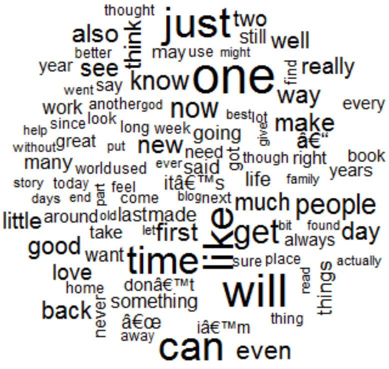
To provide a more insightful analysis, the data have been clean applying different techniques: transforming all the letters to lower case, removing the numbers and punctuation, and eliminating the stop words en English.

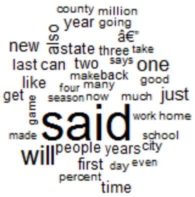
I selected a 1% sample of the data to plot the most frequent words in each file, represented in the charts below:

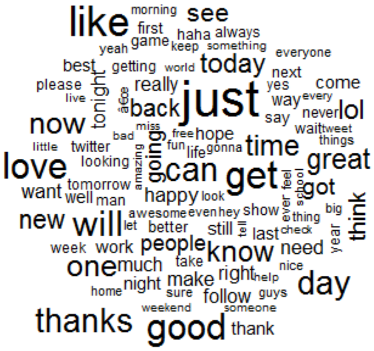

10 more common words in Blogs
0
2000
4000
6000
one
will
can
just
like
time
get
now
people
know


10 more common words in News
0
200
400
600
800
said
will
one
just
also
new
first
two
can
year


10 more common words in Twitter
0
2000
4000
6000
just
like
get
love
good
will
day
can
thanks
one







Planned method to develop the model

1. Creating the frequencies of the 2 and 3-grams. Understandin how the words link together
2. Produce an algorithm that identifies the word with the maximum probabity to appear next.

This will require further investigation on the natural language processing techniques and the functions that R offers.