Capstone Project: Milestone Report

Mohamed Taufeeq

November 12, 2014

Synopsis

The purpose of this milestone report is to explain the transformation of the raw data to tidy data including basic summary statistics, plotting, and n-gram. The source of the raw data is http://www.corpora.heliohost.org/

```
## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
##
## The following object is masked from 'package:NLP':
##
##     annotate
##
## Loading required package: qdapDictionaries
## Loading required package: qdapRegex
## Loading required package: qdapTools
## Loading required package: RColorBrewer
## WARNING: Rtools is required to build R packages, but is not currently installed.
##
## Please download and install Rtools 3.1 from http://cran.r-project.org/bin/windows/Rtools/ and then ru
##
## Attaching package: 'qdap'
##
## The following objects are masked from 'package:tm':
##
##     as.DocumentTermMatrix, as.TermDocumentMatrix
##
## The following object is masked from 'package:base':
##
##     Filter
##
## stylo version: 0.5.8.2
```

Exploratory data analysis

Basic summary statistics There are three different data file such as en_US.blogs, en_US_news, en_US_twitter in the english language data folder. These three file contains of blog posts, news articles, and twitter tweets respectively. Basic summaries of these datasets are given below:

```
##       File_name Number_of_line Number_of_word Longest_line
## 1 en_US_twitter        2360148       30683561          213
## 2    en_US_news        1010242       34798581        11384
## 3   en_US_blogs         899288       38172990        40835
```

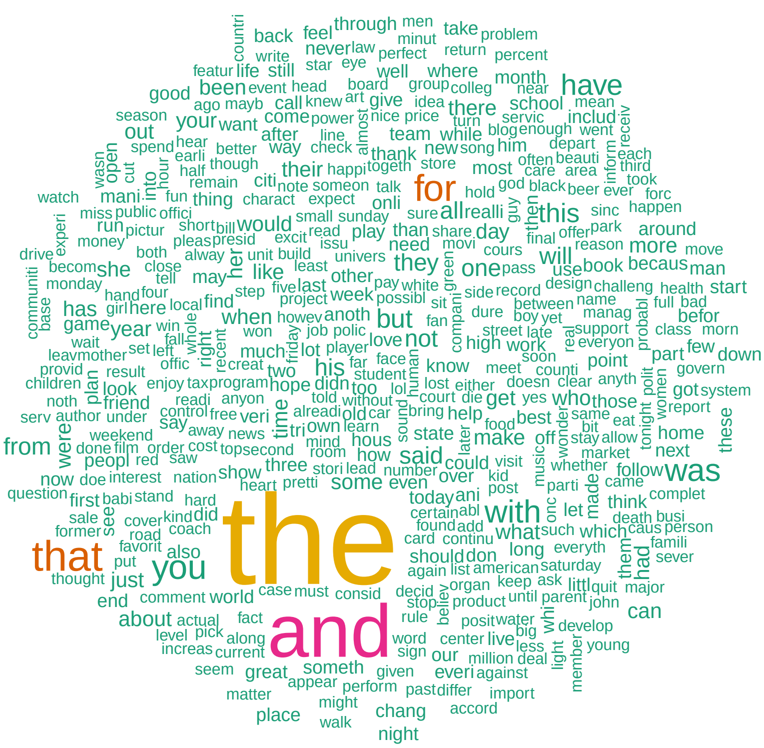
Text corpus In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts. According to the requirement of the project, build a corpus of random sample of the three data set using available R package such as tm, Rweka, wordcloud, nlp, opennlp etc. Here is an example of content of the corpus data which contains several redundent charcters such as ???.

```
## [1] "cool. i'm in w philly, but it's easy to come east/south. it looks like there are some awesome ev
## [2] "i guess they don't play basketball in montana? why on earth would you be allowed to vote???"
```

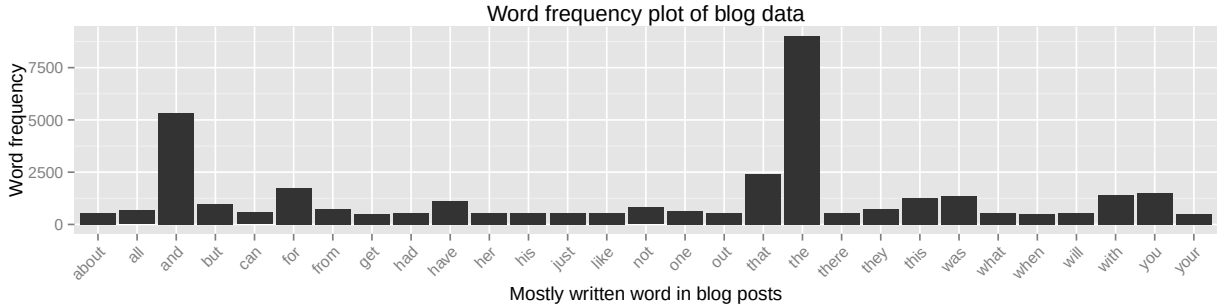
Profinity filtering Profinity filtering is basically filter out non-essential data for prediction purposes. In this phase of the exploratory analysis, perform essential repalcement of the words, sentence detection, repalcement of the non-alpha numeric characters, removing numbers and punctuation, stripping white space, and stemming words. A sample of the profinity filtered data is given below.

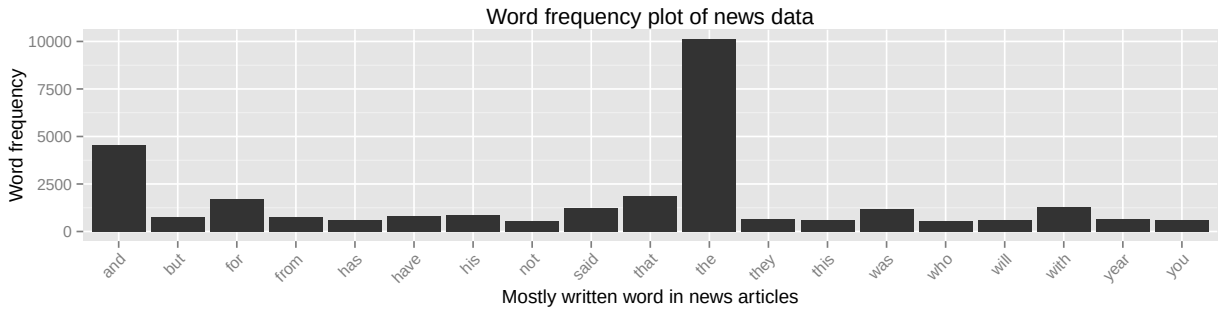
```
## [1] "cool"
## [2] "i m in w philli but it s easi to come east south"
## [3] "it look like there some awesom event"
## [4] "beer food or just beer"
## [5] "i guess they don t play basketbal in montana"
## [6] "whi on earth would you be allow to vote"
```

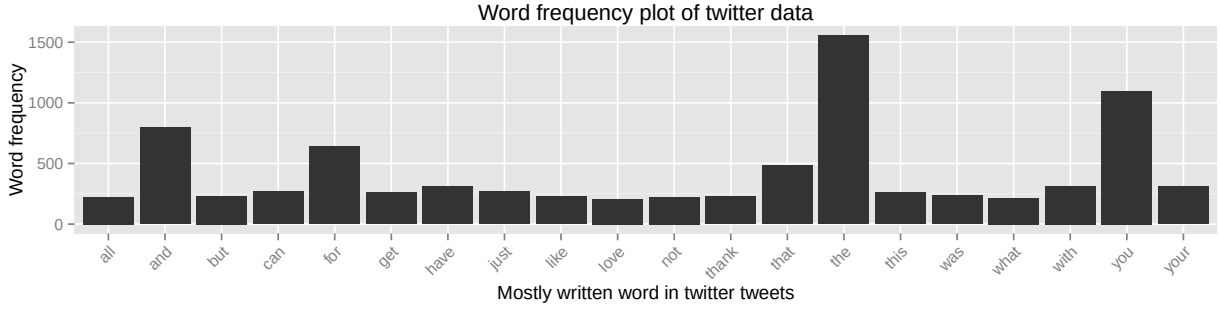
Tokenization In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. Here, filtered texts are broken into tokens using DocumentTermMatrix and nGramTokenizer function. A word cloud of the 1-gram tokenizer is provided below, and more details of n-gram tokenization is given in the very last section of the exploratory analysis.



Term frequency Several words are more frequent than all other word in the text files. Below are the three different term frequency plots of the files. All of the three plots shows which words are more common than others. X-axis of the plot represents specific words in target file and Y-axis represent corresponding frequency of the words.







From the three plots, it can say that 'the', 'and', 'you', 'for' etc. are very frequent.

Table of 1-gram, 2-gram, and 3-gram In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. Here n = 1, 2, 3. Below is shown a table of 1-gram, 2-gram, and 3-gram term, and their respected frequencies in sample data.

```
##      FTermIn1Gram Freq_1 FTermIn2Gram Freq_2   FTermIn3Gram Freq_3
## the           the  20679       of the   1969        i don t    196
## and           and  10707       in the   1759    one of the    172
## that         that   4746       to the    976       i didn t    102
## for           for   4094         it s    896        i can t     82
## you           you   3211       on the    823        i m not     74
## with         with   3010      for the    765      i want to     73
## was           was   2797        to be    698       it s not     69
## have         have   2245          i m    667     out of the     69
## this         this   2097       at the    600      you don t     69
```

```
## but          but   2000      and the    560    part of the     68
## from        from   1625        don t    544     as well as     67
## not          not   1605     with the    459     don t know     67
## they        they   1549       it was    443       go to be     62
## his          his   1458     from the    422      i ve been     60
## said        said   1446        and i    416    some of the     59
## will        will   1278        go to    406      be abl to     53
## one          one   1246      want to    382     don t have     51
## all          all   1234        i was    375    the rest of     49
## can          can   1216        it is    350      is one of     47
## out          out   1158       one of    343 the first time     47
```

Conclusion

This report covers a brief overview of the exploratory analysis of the data from taking the raw text data to building 1-gram, 2-gram, and 3-gram data for training the prediction model to predict the next unknown word.