Milestone Report

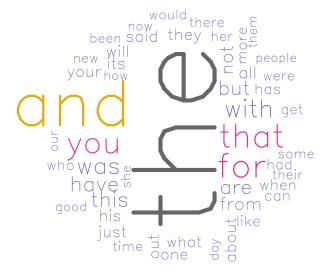
CChevalier

18 December 2015

```
# Settings
LOCALE <- "en US"
dataFolder <- "./data"
dataSampleFolder <- "./data-samples"</pre>
sampleRatio <- 1</pre>
types <- c("blogs", "news", "twitter")</pre>
set.seed(12345)
for (fileType in types) {
  preprocess(fileType, dataFolder, LOCALE, dataSampleFolder, sampleRatio )
}
## [1] "File: ./data/en_US/en_US.blogs.txt"
## [1] "
              200 Mb"
## [1] "
              37334131 words"
## [1] "
              899288 lines"
## [1] "File: ./data/en_US/en_US.news.txt"
## [1] "
           196 Mb"
## [1] "
            34372530 words"
## [1] "
             1010242 lines"
## [1] "File: ./data/en_US/en_US.twitter.txt"
          159 Mb"
## [1] "
## [1] "
              30373583 words"
## [1] "
              2360148 lines"
library(NLP)
library(tm)
library(RWeka)
library(RColorBrewer)
library(wordcloud)
corpus_sample <- Corpus(DirSource(file.path(".", dataSampleFolder, LOCALE)),</pre>
                         readerControl=list(reader=readPlain, language="en US"))
# plotWorldCloud function
#
     adapted from:
#
        Word Cloud in R
        http://www.r-bloggers.com/word-cloud-in-r/
plotWordCloud <- function(tdm, user_scale=c(3,.3)) {</pre>
 m <- as.matrix(tdm)</pre>
  v <- sort(rowSums(m), decreasing=TRUE)</pre>
  d <- data.frame(word = names(v), freq=v)</pre>
 pal <- brewer.pal(8, "Dark2")</pre>
```

```
plotWordCloud(tdm_1, c(8,.6))
title(main = "Without removing english stopwords")
```

Without removing english stopwords



```
corpus_sample_cleaned_2 <- tm_map(corpus_sample, removePunctuation)
corpus_sample_cleaned_2 <- tm_map(corpus_sample_cleaned_2, tolower)
corpus_sample_cleaned_2 <- tm_map(corpus_sample_cleaned_2,</pre>
```

```
function(x) removeWords(x, stopwords("english")))
corpus_sample_cleaned_2 <- tm_map(corpus_sample_cleaned_2, PlainTextDocument)

tdm_2 <- TermDocumentMatrix(corpus_sample_cleaned_2)

plotWordCloud(tdm_2, c(4,.3))
title(main = "With removing english stopwords")</pre>
```

With removing english stopwords work going want dont thanks

