

# Milestone Report

*CChevalier*

*18 December 2015*

```
# Settings
LOCALE <- "en_US"
dataFolder <- "./data"
dataSampleFolder <- "./data-samples"
sampleRatio <- 1
types <- c("blogs", "news", "twitter")

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"
## [1] "      159 Mb"
## [1] "      30373583 words"
## [1] "      2360148 lines"

library(NLP)
library(tm)
library(RWeka)
library(RColorBrewer)
library(wordcloud)

corpus_sample <- Corpus(DirSource(file.path(".", dataSampleFolder, LOCALE)),
  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)) {

  m <- as.matrix(tdm)
  v <- sort(rowSums(m), decreasing=TRUE)
  d <- data.frame(word = names(v), freq=v)

  pal <- brewer.pal(8, "Dark2")
```



```

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")

```

