---

title: "TWDATA"

author: "Amardeep Singh"

date: "October 3, 2018"

output:

pdf\_document: default

html\_document: default

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(error = TRUE)

```

```{r}

files <- list.files("D:\\A\_DS\\anonymisedData\\TWDATA")

fileName <- glue("D:\\A\_DS\\anonymisedData\\TWDATA\\", files[1], sep = "")

fileName <- trimws(fileName)

fileText <- glue(read\_file(fileName))

tokens <- data\_frame(text = fileText) %>% unnest\_tokens(word, text)

tokens %>%

inner\_join(get\_sentiments("bing")) %>% # pull out only sentiment words

count(sentiment) %>% # count the # of positive & negative words

spread(sentiment, n, fill = 0) %>% # made data wide rather than narrow

mutate(sentiment = positive - negative) # # of positive words - # of negative owrds

```

```{r}

SentimentAnalysis::analyzeSentiment(fileText)

```

```{r}

install.packages("installr")

library(installr)

check.for.updates.R()

install.R()

```

```{r}

install.packages("glue")

install.packages("SnowballC")

install.packages("wordcloud")

install.packages("ggplot2")

```

```{r}

install.packages("tm")

```

```{r}

slam\_url <- "https://cran.r-project.org/src/contrib/Archive/slam/slam\_0.1-37.tar.gz"

install\_url(slam\_url)

```

```{r}

install.packages('tm',dependencies = TRUE)

install.packages('NLP',dependencies = TRUE)

```

```{r}

cname <- file.path("D:\\A\_DS\\anonymisedData", "TWDATA")

cname

dir(cname)

```

```{r}

library("tm")

doc <- VCorpus(DirSource(cname))

summary(doc)

```

Length Class Mode

TWDATA.txt 2 PlainTextDocument list

<<DocumentTermMatrix (documents: 1, terms: 513812)>>

Non-/sparse entries: 513812/0

Sparsity : 0%

Maximal term length: 158

Weighting : term frequency (tf)

```{r}

install.packages("readr")

```

```{r}

inspect(docs[1])

writeLines(as.character(docs[1]))

```

```{r}

doc <- tm\_map(doc, removePunctuation)

doc <- tm\_map(doc, removeNumbers)

doc <- tm\_map(doc, tolower)

doc <- tm\_map(doc, removeWords, stopwords("english"))

doc <- tm\_map(doc, stripWhitespace)

library(SnowballC)

doc <- tm\_map(doc, stemDocument)

```

```{r}

doc <- tm\_map(doc, PlainTextDocument)

dtm<- DocumentTermMatrix(doc)

dtm

```

```{r}

length(Ccont)

length(lines)

```

```{r}

freq <- colSums(as.matrix(dtm))

length(freq)

```

```{r}

#create sort order (descending)

ord <- order(freq,decreasing=TRUE)

freq[head(ord)]

freq[tail(ord)]

```

```{r}

findFreqTerms(dtm,lowfreq=80)

```

```{r}

dtm <- TermDocumentMatrix(doc)

m <- as.matrix(dtm)

v <- sort(rowSums(m),decreasing=TRUE)

d <- data.frame(word = names(v),freq=v)

d

```

```{r}

library(wordcloud)

wordcloud(names(d),freq,min.freq=10)

```

The same world cloud is created but with color formats for easier differentiation and interpretation.

```{r}

set.seed(1234)

wordcloud(words = d$word, freq = d$freq, min.freq = 100,

max.words=200, random.order=FALSE, rot.per=0.35,

colors=brewer.pal(8, "Dark2"))

```

