# MongoTwitterReport

*David Churchman, Lakshmi Bobbillapati*

*March 31, 2018*

## Part 1: Download Twitter Data to MongoDB Database

First we created a Twitter API token and used it download a large set of tweets about data science.

```
library(mongolite)
library(twitteR)
library(stringi)
library(ROAuth)
library(tm)
```

```
## Loading required package: NLP
```

```
consumer_key <- '####'
consumer_secret <- '####'
access_token <- '####'
access_secret <- '####'
#I saved my secrets in APIcodes.R
source('APIcodes.R')
## Twitter authentication
setup_twitter_oauth(consumer_key, consumer_secret, access_token,
                    access_secret)
```

```
## [1] "Using direct authentication"
```

```
#Search twitter, restrict to English.
tweets  <- searchTwitter('#datascience',n=10000, since='2018-03-25',until='2018-03-28', lang='en')
# convert tweets to a data frame
tweets.df <- twListToDF(tweets)

#Limit to more interesting columns
tweetsfew <- tweets.df[c('text','favoriteCount','created','screenName','retweetCount','isRetweet')]


#MongoDB didn't like something about the encoding of tweets, so changing from mostly UTF8 to ASCII
tweetsfew$text <- stri_enc_toascii(tweetsfew$text)

#Create MongoDB database
#Make sure MongoDB installed and execute mongod app
collection = mongo(collection = "tweets2", db = "datatweets") # create connection, database and collect
collection$insert(tweetsfew)
```

```
## List of 5
##  $ nInserted  : num 10000
##  $ nMatched   : num 0
##  $ nRemoved   : num 0
##  $ nUpserted  : num 0
##  $ writeErrors: list()
```

```r
collection$count()
```

```
## [1] 50100
```

```r
collection$iterate()$one()
```

```
## $text
## [1] "RT @IainLJBrown: HP Introduces World's Most Powerful Workstation for Machine Learning Developmen
##
## $favoriteCount
## [1] 0
##
## $created
## [1] "2018-03-27 18:59:51 CDT"
##
## $screenName
## [1] "NkrumaIgnatov"
##
## $retweetCount
## [1] 31
##
## $isRetweet
## [1] TRUE
```

```r
#MongoDB didn't like something about the encoding of tweets, so changing from mostly UTF8 to ASCII
tweetsfew$text <- suppressWarnings(stri_enc_toascii(tweetsfew$text))
```

## Part 2: Data clean up

```r
# build a corpus, and specify the source to be character vectors
myCorpus <- Corpus(VectorSource(tweetsfew$text))

# convert to lower case
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
# remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
# remove anything other than English letters or space
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
# remove stopwords #Play around with these later!!!######################
myStopwords <- c(stopwords('english'),"datascience", "via","iainljbrown","rt","data")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
# remove extra whitespace
myCorpus <- tm_map(myCorpus, stripWhitespace)

tdm <- TermDocumentMatrix(myCorpus, control = list(wordLengths = c(1, Inf)))

# inspect frequent words
freq.terms <- findFreqTerms(tdm, lowfreq = 1000)
term.freq <- rowSums(as.matrix(tdm))
term.freq <- subset(term.freq, term.freq >= 1000)
df <- data.frame(term = names(term.freq), freq = term.freq)
```

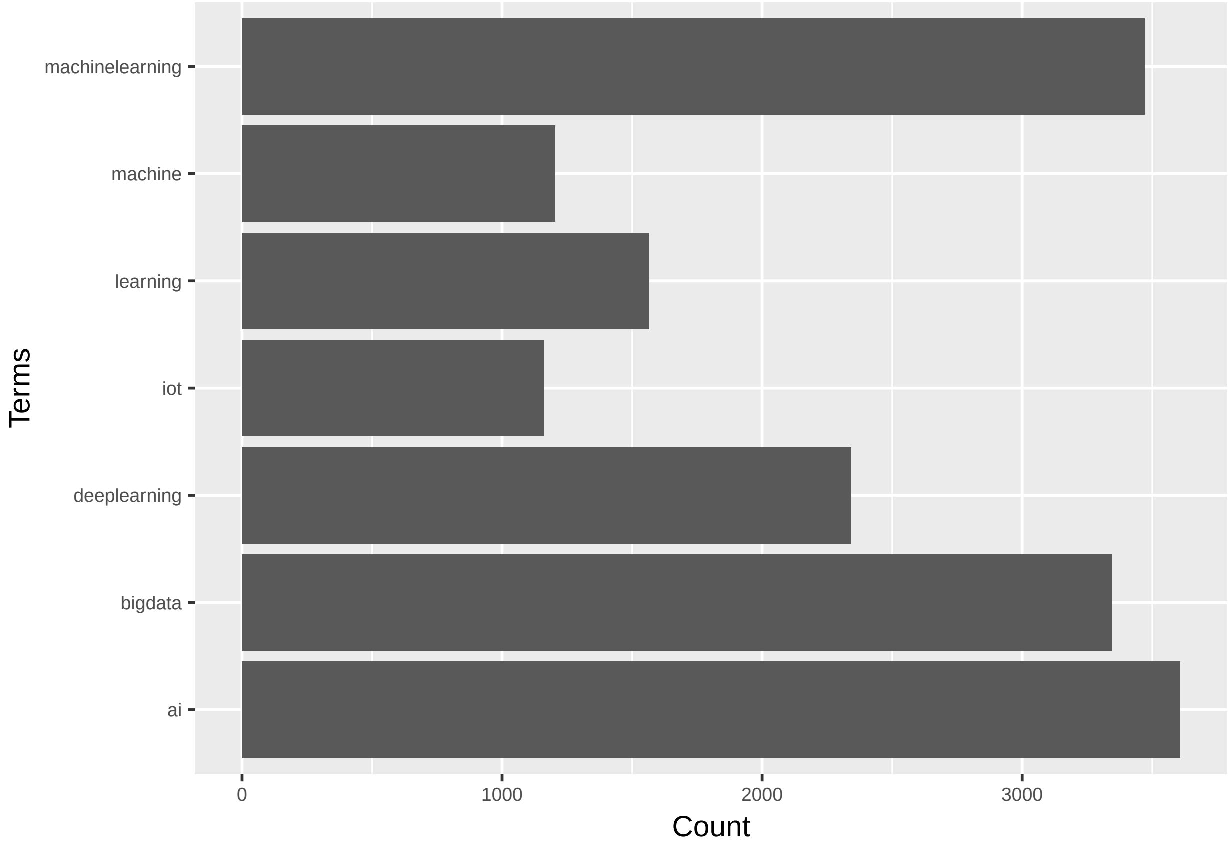
## Part 3: Text analysis of tweets

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate
```

```
ggplot(df, aes(x=term, y=freq)) + geom_bar(stat="identity") +
  xlab("Terms") + ylab("Count") + coord_flip() +
  theme(axis.text=element_text(size=7))
```

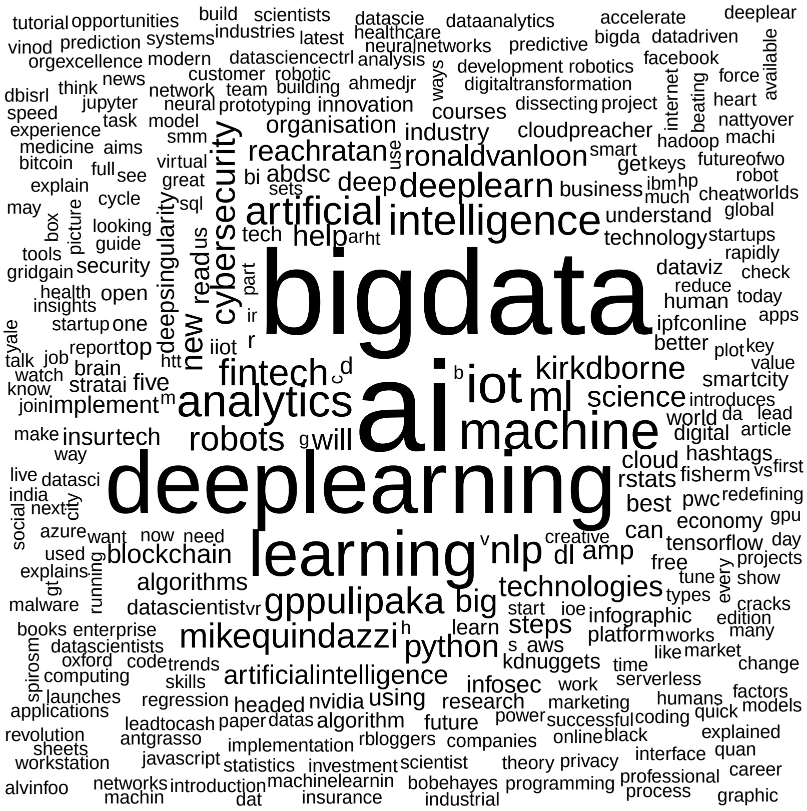


```
m <- as.matrix(tdm)
# calculate the frequency of words and sort it by frequency
word.freq <- sort(rowSums(m), decreasing = T)

# plot word cloud
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
wordcloud(words = names(word.freq), freq = word.freq, min.freq = 50,
          random.order = F)
```



## Sources

https://datascienceplus.com/using-mongodb-with-r/ http://www.rdatamining.com/docs/twitter-analysis-with-r