Project_3

Robert Sellers

March 15, 2016

Libraries used.

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.4
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(stringr)
library(twitteR)
```

```
## Warning: package 'twitteR' was built under R version 3.2.4

##
## Attaching package: 'twitteR'

## The following objects are masked from 'package:dplyr':
##
##     id, location
```

```
require(wordcloud)
```

```
## Loading required package: wordcloud
```

```
## Warning: package 'wordcloud' was built under R version 3.2.4
```

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

```
require(tm)
```

```
## Loading required package: tm
```

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

Twitter OAuth. Will not be evaluated. This is only required when generating new datasets. Both included, only one needed.

```
#Robert's Credentials
#consumer_key <- "DaAA9z8QvnxsdL0SpIr1oYwvP"
#consumer_secret <- "bCfsuODQyoYKMxPoHhZy2LxvvVqBvSM1LemzBtqm6YFeylWKUE"
#access_token <-    "558596891-tDxN7T34cyVJJaBc4ExGTAq6wRfFBBlyHb2IzQvM"
#access_secret <-"nHqi5sVT2XRoCoSu0dYQnboCg1h35w5hRvtg657t8ROX8"
#setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

#Chirag's credentials
#consumer_key <- "DaAA9z8QvnxsdL0SpIr1oYwvP"
#consumer_secret <- "bCfsuODQyoYKMxPoHhZy2LxvvVqBvSM1LemzBtqm6YFeylWKUE"
#access_token <-     "558596891-tDxN7T34cyVJJaBc4ExGTAq6wRfFBBlyHb2IzQvM"
#access_secret <-"nHqi5sVT2XRoCoSu0dYQnboCg1h35w5hRvtg657t8ROX8"
```

Sample data mining code. The following code was used to generate the .csv files that you will find inside: https://github.com/RobertSellers/SlackProjects/tree/master/data

```
twitter_results_Feb_March_19<-searchTwitter("#datascience", n=10000)
twitter_results_Feb_March_19 <- Map(as.data.frame, twitter_results_Feb_March_19 )
twitter_results_Feb_March_19  <- rbindlist(twitter_results_Feb_March_19  )
write.csv(twitter_results_Feb_March_19 , file = "C:/Users/Robert/Desktop/CUNY/GitHub/R/data/twitter_resu
```

Loading the data sources. Currently dating from 3/16, 3/18, 3/19. Will run only 3/16 for this.

```
twitter_results_march_16<-read.csv(file="https://raw.githubusercontent.com/RobertSellers/SlackProjects/m
```

```
twitter_results_march_18<-read.csv(file="https://raw.githubusercontent.com/RobertSellers/SlackProjects/m
twitter_results_march_19<-read.csv(file="https://raw.githubusercontent.com/RobertSellers/SlackProjects/m
```

Word Cloud function.

To Do: We may want a second "stopWords" variable input.

To Do: The results from this word cloud (that are relevant) should be added to the skills.csv.

To Do: Find a way to export a frequency table from this data?

```
dataScienceWordCloud<-function(twitterData){
  text<-twitterData$text
  corpus<- Corpus(VectorSource(text))
  corpus <- tm_map(corpus, stripWhitespace)
  corpus <- tm_map(corpus, content_transformer(tolower))
  corpus <- tm_map(corpus, removeNumbers)

  toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
  corpus <- tm_map(corpus, toSpace, "http\\S+\\s*")
  corpus <- tm_map(corpus, toSpace, "http\\S+\\s*") #twice????
  corpus <- tm_map(corpus, toSpace, "#")
  corpus <- tm_map(corpus, toSpace, "http\\w+")
  corpus <- tm_map(corpus, toSpace, "@\\w+")

  corpus <- tm_map(corpus, toSpace, "https\\w+")
  corpus <- tm_map(corpus, toSpace, "uselect")
  corpus <-tm_map(corpus, removeWords, c(stopwords("english"),"#"))

  wordcloud(corpus,random.order=F,min.freq=1,max.words=400,colors=brewer.pal(8, "Dark2"))
}
```

Running the word cloud on March 16

```
dataScienceWordCloud(twitter_results_march_16)
```

Lookup table function.

ToDo: This uses the "skill.csv" as a look up table that is also applied to the other team's URL method. We ought to add new values to that csv based on what we find in the twitter dataset.

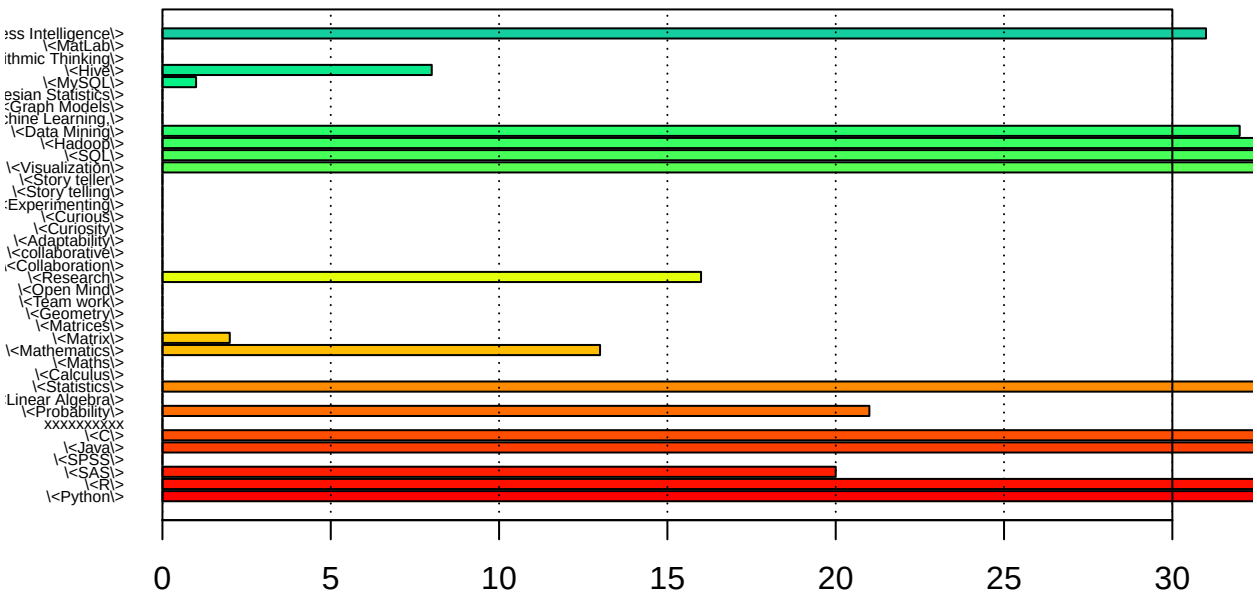
```
lutSkills<-read.csv(file= "https://raw.githubusercontent.com/RobertSellers/SlackProjects/master/data/ski

lookupFrequencies <-function(twitterData,lookupTable) {
  lookupTable<-as.data.frame(lookupTable)
  lookupTable$Skill<-paste0("\\<",lookupTable$Skill,"\\>")
  lookupTable["counts"]<-NA
  lookupTable$Skill[7] <- "xxxxxxxxxx" #C++ not working as a keyword
  i<-1
  for(i in 1:nrow(lookupTable)) {
      lookupTable$counts[i]<-length(grep(lookupTable$Skill[i], twitterData$text))
  }
  colfunc<-colorRampPalette(c("red","yellow","springgreen","royalblue"))
  colnamesbarplot <- as.character(lookupTable$Skill)
  barplot(lookupTable$counts,main="Word Counts",horiz=TRUE,col=(colfunc(50)),axes=TRUE, names.arg=colnar
grid(nx=NULL, ny=NA,col="black")
box()
}
```

Running the function on March 16th

To Do - Fix plot & refine the function. Sort the data and continue to update the skills lookup table to get better data.

To Do - We ultimately want to run this on each date to ensure that we have consistency between dates and then ultimately to combine all of the data.

```
lookupFrequencies(twitter_results_march_16,lutSkills)
```


Word Counts
ess Intelligence\>
\<MatLab\>
ithmic Thinking\>
\<Hive\>
\<MySQL\>
esian Statistics\>
<Graph Models\>
hine Learning \>
\<Data Mining\>
\<Hadoop\>
\<SQL\>
\<Visualization\>
\<Story teller\>
\<Story telling\>
<Experimenting\>
\<Curious\>
\<Curiosity\>
\<Adaptability\>
\<collaborative\>
<Collaboration\>
\<Research\>
\<Open Mind\>
\<Team work\>
\<Geometry\>
\<Matrices\>
\<Matrix\>
\<Mathematics\>
\<Maths\>
\<Calculus\>
\<Statistics\>
<Linear Algebra\>
\<Probability\>
xxxxxxxxxx
\<C\>
\<Java\>
\<SPSS\>
\<SAS\>
\<R\>
\<Python\>
0
5
10
15
20
25
30