Final Project

```
clinton = read.csv("clinton_sentiment.csv", stringsAsFactors = FALSE)
trump = read.csv("trump_sentiment.csv", stringsAsFactors = FALSE)
head(clinton)
```

```
##   X Sentiment_Score
## 1 0              -1
## 2 1               1
## 3 2               1
## 4 3               2
## 5 4               0
## 6 5               0
##
Tweet
## 1 @AndreaTantaros @LeahR77 Hillary Clinton must have her reasons!  And, reasons are like excuses!  But, her
explanation would be in-lightning!
## 2                    In that whole #GOPDebate, Hillary Clinton got 39 mentions and Bernie Sanders got zero.
Talk about counting your chickens.
## 3                    In that whole #GOPDebate, Hillary Clinton got 39 mentions and Bernie Sanders got zero.
Talk about counting your chickens.
## 4 "The ridiculous left-wing 'crusade' against Hillary Clinton needs to "stop:" https://t.co/4xZ9RirHRO #Hil
laryClinton #Hilary #Hilary2016 #US
## 5                                                                                    Lol why does Fiorina keep bri
nging Hillary Clinton up? #PlsChill
## 6                        Dans le NYTimes, Hillary Clinton veut une action coordonnée des acteurs du net co
ntre ISIS - https://t.co/Jcck3qxEgO
```

```
str(clinton)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Sentiment_Score: int  -1 1 1 2 0 0 0 0 0 -1 ...
##  $ Tweet          : chr  "@AndreaTantaros @LeahR77 Hillary Clinton must have her reasons!  And, reasons are
like excuses!  But, her explanation would be "| __truncated__ "In that whole #GOPDebate, Hillary Clinton got 3
9 mentions and Bernie Sanders got zero. Talk about counting your chickens." "In that whole #GOPDebate, Hillary
Clinton got 39 mentions and Bernie Sanders got zero. Talk about counting your chickens." "\"The ridiculous lef
t-wing 'crusade' against Hillary Clinton needs to \"stop:\" https://t.co/4xZ9RirHRO #HillaryClinton #Hilary"|
__truncated__ ...
```

```
head(trump)
```

```
##   X Sentiment_Score
## 1 0              -4
## 2 1              -5
## 3 2               0
## 4 3               3
## 5 4               1
## 6 5              -1
##
Tweet
## 1      RT @tyriquex: Bernie Sanders explaining what's so dangerous about Donald Trump running for president
. https://t.co/W7PnfbFduN
## 2 RT @ForQ2: .@hardball_chris @lonepatrick Why didn't you ask donald trump what is Hawaii investigators fou
nd? ? trump is repugnant!
## 3                                                                                    RT @muuugn: Donald Trump Says "China
" https://t.co/yfhP7UJSd8
## 4                RT @lovelive_txt: who would be a better president\nrt for nico yazawa\nfav for donald tru
mp http://t.co/QrtNcqTL3J
## 5                                    #RealDonaldTrump is the greatest business mind: https://t.co/jFeDBWohv
b https://t.co/uLK7ZV6Ns5
## 6                                                                                    @klustout Russia, Pakistan, Nort
h Korea and Donald Trump.
```

```
str(trump)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
```

```
##  $ Sentiment_Score: int  -4 -5 0 3 1 -1 0 0 2 -4 ...
##  $ Tweet          : chr  "RT @tyriquex: Bernie Sanders explaining what's so dangerous about Donald Trump ru
nning for president. https://t.co/W7PnfbFduN" "RT @ForQ2: .@hardball_chris @lonepatrick Why didn't you ask don
ald trump what is Hawaii investigators found? ? trump is repugna"| __truncated__ "RT @muuugn: Donald Trump Say
s \"China\" https://t.co/yfhP7UJSd8" "RT @lovelive_txt: who would be a better president\nrt for nico yazawa\nfa
v for donald trump http://t.co/QrtNcqTL3J" ...
```

We are reading the data into R and using str() and head() to look at it.

```
clinton$Negative = as.factor(clinton$Sentiment_Score < 0)
table(clinton$Negative)
```

```
##
## FALSE  TRUE
##   680   320
```

```
trump$Negative = as.factor(trump$Sentiment_Score < 0)
table(trump$Negative)
```

```
##
## FALSE  TRUE
##   606   394
```

```
clinton$Positive = as.factor(clinton$Sentiment_Score > 0)
table(clinton$Positive)
```

```
##
## FALSE  TRUE
##   752   248
```

```
trump$Positive = as.factor(trump$Sentiment_Score > 0)
table(trump$Positive)
```

```
##
## FALSE  TRUE
##   619   381
```

I added a negative variable to both clinton and trump dataset if the sentiment score was less than 0. I will use this later in the decision tree.

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.2.2
```

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

```
## Warning: package 'NLP' was built under R version 3.2.2
```

```
library(SnowballC)
```

```
## Warning: package 'SnowballC' was built under R version 3.2.3
```

```
corpus_clinton = Corpus(VectorSource(clinton$Tweet))
corpus_trump = Corpus(VectorSource(trump$Tweet))
corpus_clinton = tm_map(corpus_clinton, tolower)
corpus_trump = tm_map(corpus_trump, tolower)
corpus_clinton = tm_map(corpus_clinton,removePunctuation)
corpus_trump = tm_map(corpus_trump,removePunctuation)
corpus_clinton = tm_map(corpus_clinton, removeWords, c("hillary clinton", stopwords("english")))
corpus_trump = tm_map(corpus_trump, removeWords, c("donald trump", stopwords("english")))
corpus_clinton = tm_map(corpus_clinton, stemDocument)
corpus_trump = tm_map(corpus_trump, stemDocument)
corpus_clinton <- tm_map(corpus_clinton, PlainTextDocument)
corpus_trump <- tm_map(corpus_trump, PlainTextDocument)
```

Above I am cleaning the data cusing tm_map by putting the tweets in lower case, removing punctuation, removing stop words and steming the data.

```
frequencies_clinton = DocumentTermMatrix(corpus_clinton)
frequencies_trump = DocumentTermMatrix(corpus_trump)
frequencies_trump
```

```
## <<DocumentTermMatrix (documents: 1000, terms: 1641)>>
## Non-/sparse entries: 8714/1632286
## Sparsity           : 99%
## Maximal term length: 24
## Weighting          : term frequency (tf)
```

```
frequencies_clinton
```

```
## <<DocumentTermMatrix (documents: 1000, terms: 614)>>
## Non-/sparse entries: 9580/604420
## Sparsity           : 98%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

```
findFreqTerms(frequencies_clinton,lowfreq=40)
```

```
##  [1] "<U+0097>govhowarddean"     "bernie"             "call"
##  [4] "campaign"           "clintons"           "dominance"
##  [7] "donald"             "ever"               "families"
## [10] "gopdebate"          "got"                "hillary"
## [13] "hillaryclinton"     "hillarys"           "http<U+0085>"
## [16] "https<U+0085>"             "httpstcogf2y8pkeii" "httpstcounoqzfjzck"
## [19] "httpstcozh2tfftkxp" "ibd"                "ibdeditorials"
## [22] "isis"               "journalists"        "kids"
## [25] "latest"             "like"               "martyaramirez"
## [28] "matched"            "moms"               "nearly"
## [31] "new"                "news"               "obama"
## [34] "obamaplus"          "poll"               "running"
## [37] "shes"               "solar"              "strategy"
## [40] "time"               "trump"              "trump<U+0092>s"
## [43] "via"                "video"              "washingtonpost"
## [46] "will"               "youtub"
```

```
findFreqTerms(frequencies_trump,lowfreq=40)
```

```
##  [1] "ban"                "bernie"             "bush"
##  [4] "business"           "call"               "dangerous"
##  [7] "debate"             "donald"             "explaining"
## [10] "full"               "gopdebate"          "greatest"
## [13] "hatred"             "httpstco<U+0085>"          "httpstcojfedbwohvb"
## [16] "httpstcow7pnfbfdun" "jeb"                "just"
## [19] "like"               "malala"             "mind"
## [22] "muslim"             "muslims"            "people"
## [25] "president"          "realdonaldtrump"    "republican"
## [28] "running"            "sanders"            "says"
## [31] "tragic"             "trump"              "trumps"
## [34] "tyriquex"           "views"              "whats"
## [37] "yousafzai"
```

```
sparse_trump = removeSparseTerms(frequencies_trump,0.995)
sparse_clinton = removeSparseTerms(frequencies_clinton,0.995)
tweetsSparse_clinton = as.data.frame(as.matrix(sparse_clinton))
tweetsSparse_trump = as.data.frame(as.matrix(sparse_trump))
colnames(tweetsSparse_trump) = make.names(colnames(tweetsSparse_trump))
colnames(tweetsSparse_clinton) = make.names(colnames(tweetsSparse_clinton))
```

Above we are doing some analysis to the data by seeing which words occur with a frequence of 40 times or more and removing “sparse” words.

/nl library(rpart)
/nl library(rpart.plot) /nl tweetCART_clinton = rpart(Negative ~ ., data = tweetsSparse_trump, method = “class”) /nl tweetCART_trump = rpart(Negative ~ ., data = tweetsSparse_clinton, method = “class”) /nl prp(tweetCART_clinton) /nl prp(tweetCART_trump)

The above 6 lines of code runs fine in my R console but I could not get it to output properly in R markdown. R markdown compacted it in 2 lines as well during the R markdown output. I tried to make the line breaks more clear by adding /nl in the R markdown.

Above we are building a CART model for negative sentiment for the two candidates. So the graph should give some breakdown of if a certain

word is said, then it decreases or increases the chance of the tweet to be negative in sentiment. For Donald Trump, it is interesting that when the word “realdona” is mentioned, that decreases the chance of the tweet to have a negative sentiment. That word is stemmed, so the real word is “realdonaldtrump” which is Donald Trump’s twitter handle.

The whole analysis was guided by the text analytics section of the MOOC analytics edge by edx.org. https://www.edx.org/course/analytics-edge-mitx-15-071x-0