HW6

Part1.

```
library("dplyr")
library("tidyr")
library("ggplot2")
library("tm")           # Text mining package
library("wordcloud2")   # Package for building word clouds
library("syuzhet")      # Package for sentement analysis
library("stringr")      # Package for work with strings
library("class")        # KNN
library("e1071")        # For SVM
library("igraph")
library("SnowballC")
library("wordcloud")
library("randomForest")

library("readr")
```

```
Apple1 <- read_csv("/home/nesma/SemesterII/BusinessDataAnalytics/HW6/Apple1.csv")
head(Apple1, 3)
```

```
## # A tibble: 3 x 4
##   text                                created                  id sentiment
##   <chr>                               <dttm>                <dbl> <chr>
## 1 RT @option_snipper: $AAPL beat on ~ 2017-08-01 20:31:56 8.92e17 positive
## 2 RT @option_snipper: $AAPL beat on ~ 2017-08-01 20:31:55 8.92e17 positive
## 3 Let's see this break all timers. $~ 2017-08-01 20:31:55 8.92e17 neutral
```

```
str(Apple1)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1000 obs. of  4 variables:
##  $ text     : chr  "RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B,
##  $ created  : POSIXct, format: "2017-08-01 20:31:56" "2017-08-01 20:31:55" ...
##  $ id       : num  8.92e+17 8.92e+17 8.92e+17 8.92e+17 8.92e+17 ...
##  $ sentiment: chr  "positive" "positive" "neutral" "negative" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   text = col_character(),
##   ..   created = col_datetime(format = ""),
##   ..   id = col_double(),
##   ..   sentiment = col_character()
##   .. )
```

```
# change encoding of our texts to "UTF-8"
Apple1Text <- iconv(Apple1$text, to = "utf-8")
# converting character vectors to specified encodings
corpus1 <- Corpus(VectorSource(Apple1Text))
#View Corpus
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
```

```
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B htt
## [2] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B htt
## [3] Let's see this break all timers. $AAPL 156.89
## [4] RT @SylvaCap: Things might get ugly for $aapl with the iphone delay. With $aapl down that means a
## [5] $AAPL - wow! This was supposed to be a throw-away quarter and AAPL beats by over 500 million in r
```

```
#Remove URL

removeURL <- function(x) gsub('https://[[:alnum:]|[:punct:]]*', '', x)
corpus1 <- tm_map(corpus1, content_transformer(removeURL))
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [2] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [3] Let's see this break all timers. $AAPL 156.89
## [4] RT @SylvaCap: Things might get ugly for $aapl with the iphone delay. With $aapl down that means a
## [5] $AAPL - wow! This was supposed to be a throw-away quarter and AAPL beats by over 500 million in r
```

```
#Remove @
removeat <- function(x) gsub("@\\w+ *", "",x )
corpus1 <- tm_map(corpus1, content_transformer(removeat))
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT : $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [2] RT : $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [3] Let's see this break all timers. $AAPL 156.89
## [4] RT : Things might get ugly for $aapl with the iphone delay. With $aapl down that means almost all
## [5] $AAPL - wow! This was supposed to be a throw-away quarter and AAPL beats by over 500 million in r
```

```
#Remove Dollar Sign
removedollar <- function(x) gsub("\\$\\w+ *", "",x )
corpus1 <- tm_map(corpus1, content_transformer(removedollar))
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT : beat on both eps and revenues. SEES 4Q REV. -, EST. .1B
## [2] RT : beat on both eps and revenues. SEES 4Q REV. -, EST. .1B
## [3] Let's see this break all timers. 156.89
## [4] RT : Things might get ugly for with the iphone delay. With down that means almost all of the FANG
## [5] - wow! This was supposed to be a throw-away quarter and AAPL beats by over 500 million in revenue
```

```
#Remove Punctuation
corpus1 <- tm_map(corpus1, removePunctuation)
```

```r
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  beat on both eps and revenues SEES 4Q REV  EST 1B
## [2] RT  beat on both eps and revenues SEES 4Q REV  EST 1B
## [3] Lets see this break all timers 15689
## [4] RT  Things might get ugly for with the iphone delay With down that means almost all of the FANG s
## [5]  wow This was supposed to be a throwaway quarter and AAPL beats by over 500 million in revenue Tr
```

```r
#Remove Numbers
corpus1 <- tm_map(corpus1, removeNumbers)
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  beat on both eps and revenues SEES Q REV  EST B
## [2] RT  beat on both eps and revenues SEES Q REV  EST B
## [3] Lets see this break all timers
## [4] RT  Things might get ugly for with the iphone delay With down that means almost all of the FANG s
## [5]  wow This was supposed to be a throwaway quarter and AAPL beats by over  million in revenue Trill
```

```r
#to lowercase
corpus1 <- tm_map(corpus1, content_transformer(tolower))
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt  beat on both eps and revenues sees q rev  est b
## [2] rt  beat on both eps and revenues sees q rev  est b
## [3] lets see this break all timers
## [4] rt  things might get ugly for with the iphone delay with down that means almost all of the fang s
## [5]  wow this was supposed to be a throwaway quarter and aapl beats by over  million in revenue trill
```

```r
#Remove stop words
corpus1 <- tm_map(corpus1, removeWords, stopwords('english'))
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt  beat   eps  revenues sees q rev  est b
## [2] rt  beat   eps  revenues sees q rev  est b
## [3] lets see  break  timers
## [4] rt  things might get ugly    iphone delay    means almost    fang stocks   pos...
## [5]  wow   supposed    throwaway quarter  aapl beats    million  revenue trillion dollar company
```

```r
#Remove White spaces
corpus1 <- tm_map(corpus1, stripWhitespace)
inspect(corpus1[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt beat eps revenues sees q rev est b
## [2] rt beat eps revenues sees q rev est b
## [3] lets see break timers
## [4] rt things might get ugly iphone delay means almost fang stocks pos...
## [5]  wow supposed throwaway quarter aapl beats million revenue trillion dollar company
```

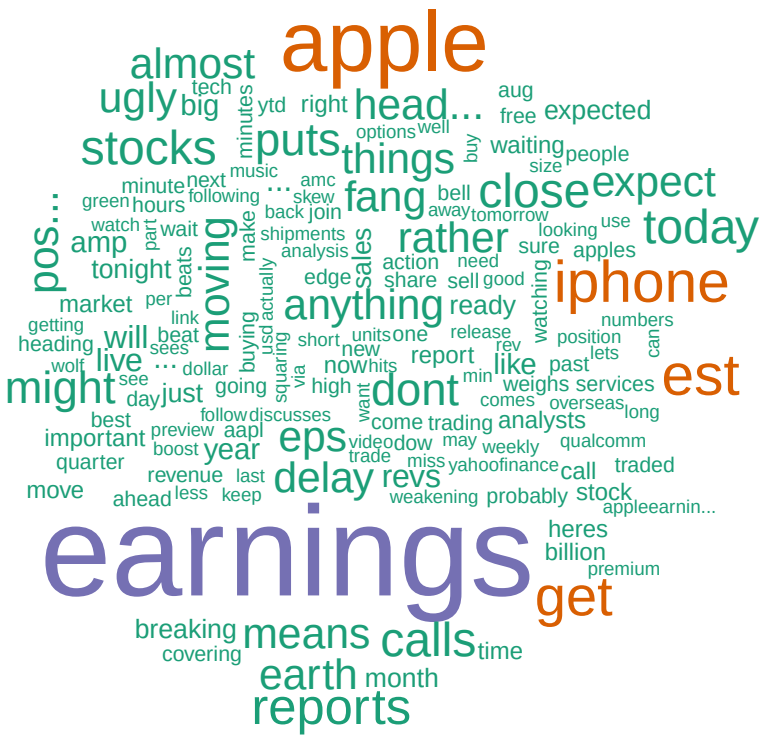
```
dtm1 <- DocumentTermMatrix(corpus1)
inspect(dtm1)
```

```
## <<DocumentTermMatrix (documents: 1000, terms: 1315)>>
## Non-/sparse entries: 7164/1307836
## Sparsity           : 99%
## Maximal term length: 22
## Weighting          : term frequency (tf)
## Sample             :
##      Terms
## Docs  apple calls close earnings est get iphone reports stocks today
##   428     0     0     0        1   0   0      0       0      0     1
##   488     0     0     0        0   0   0      0       0      0     0
##   495     0     1     0        1   0   0      0       0      0     0
##   523     0     0     0        0   0   0      0       0      0     0
##   537     0     0     0        1   0   0      0       0      0     0
##   600     0     0     0        0   0   0      0       0      0     0
##   607     0     0     0        0   0   0      0       0      0     0
##   612     0     0     0        0   0   0      0       0      0     0
##   743     0     0     0        1   0   0      0       0      1     0
##   807     0     0     0        1   0   0      0       0      0     0
```

```
dtm1 <- as.data.frame(as.matrix(dtm1))
freq1 = data.frame(sort(colSums(as.matrix(dtm1)), decreasing=TRUE))
#top 2 frequent words:
head(freq1,2)
```

```
##          sort.colSums.as.matrix.dtm1....decreasing...TRUE.
## earnings                                                362
## apple                                                   224
```

```
wordcloud(rownames(freq1), freq1[,1], min.freq = 8, colors=brewer.pal(1, "Dark2"))
```




```
#Work for the second dataset

Apple2 <- read_csv("/home/nesma/SemesterII/BusinessDataAnalytics/HW6/Apple2.csv")
head(Apple2)
```

```
## # A tibble: 6 x 4
##   text                                created                  id sentiment
##   <chr>                               <dttm>                <dbl> <chr>
## 1 RT @philstockworld: Whipsaw Wednes~ 2017-08-02 14:25:07 8.93e17 neutral
## 2 RT @philstockworld: Whipsaw Wednes~ 2017-08-02 14:25:04 8.93e17 neutral
## 3 RT @stockpicklist: $NBDR is making~ 2017-08-02 14:25:03 8.93e17 neutral
## 4 RT @philstockworld: Whipsaw Wednes~ 2017-08-02 14:25:00 8.93e17 neutral
## 5 RT @TDAJJKinahan: With $AAPL behin~ 2017-08-02 14:24:56 8.93e17 positive
## 6 $AAPL Im still waiting.             2017-08-02 14:24:54 8.93e17 neutral
```

```
str(Apple2)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1000 obs. of  4 variables:
##  $ text     : chr  "RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $
##  $ created  : POSIXct, format: "2017-08-02 14:25:07" "2017-08-02 14:25:04" ...
##  $ id       : num  8.93e+17 8.93e+17 8.93e+17 8.93e+17 8.93e+17 ...
##  $ sentiment: chr  "neutral" "neutral" "neutral" "neutral" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   text = col_character(),
##   ..   created = col_datetime(format = ""),
##   ..   id = col_double(),
```

```
##   ..   sentiment = col_character()
##   .. )
```

```
# change encoding of our texts to "UTF-8"
Apple2Text <- iconv(Apple2$text, to = "utf-8")
# converting character vectors to specified encodings
corpus2 <- Corpus(VectorSource(Apple2Text))
#View Corpus
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [2] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [3] RT @stockpicklist: $NBDR is making moves toward #triple #digit #gains Make your move #today #Sto
## [4] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [5] RT @TDAJJKinahan: With $AAPL behind us, focus turns to TSLA, with hopes of more insight into the
```

```
#Remove URL
corpus2 <- tm_map(corpus2, content_transformer(removeURL))
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [2] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [3] RT @stockpicklist: $NBDR is making moves toward #triple #digit #gains Make your move #today #Sto
## [4] RT @philstockworld: Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #
## [5] RT @TDAJJKinahan: With $AAPL behind us, focus turns to TSLA, with hopes of more insight into the
```

```
#remove at
corpus2 <- tm_map(corpus2, content_transformer(removeat))
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #Dollar --
## [2] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #Dollar --
## [3] RT : $NBDR is making moves toward #triple #digit #gains Make your move #today #StockMarket #paid
## [4] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 $DIA $DXD $AAPL #Oil #Dollar --
## [5] RT : With $AAPL behind us, focus turns to TSLA, with hopes of more insight into the company's del
```

```
#Remove Dollar Sign
corpus2 <- tm_map(corpus2, content_transformer(removedollar))
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 #Oil #Dollar --
```

```
## [2] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 #Oil #Dollar --
## [3] RT : is making moves toward #triple #digit #gains Make your move #today #StockMarket #paid ...
## [4] RT : Whipsaw Wednesday - The View from Dow 22,000 #Dow22000 #Oil #Dollar --
## [5] RT : With behind us, focus turns to TSLA, with hopes of more insight into the company's delivery
```

```
#Remove Punctuation
corpus2 <- tm_map(corpus2, removePunctuation)
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  Whipsaw Wednesday  The View from Dow 22000 Dow22000 Oil Dollar
## [2] RT  Whipsaw Wednesday  The View from Dow 22000 Dow22000 Oil Dollar
## [3] RT  is making moves toward triple digit gains Make your move today StockMarket paid ...
## [4] RT  Whipsaw Wednesday  The View from Dow 22000 Dow22000 Oil Dollar
## [5] RT  With behind us focus turns to TSLA with hopes of more insight into the company's delivery sc
```

```
#Remove Numbers
corpus2 <- tm_map(corpus2, removeNumbers)
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  Whipsaw Wednesday  The View from Dow  Dow Oil Dollar
## [2] RT  Whipsaw Wednesday  The View from Dow  Dow Oil Dollar
## [3] RT  is making moves toward triple digit gains Make your move today StockMarket paid ...
## [4] RT  Whipsaw Wednesday  The View from Dow  Dow Oil Dollar
## [5] RT  With behind us focus turns to TSLA with hopes of more insight into the company's delivery sc
```

```
#to lowercase
corpus2 <- tm_map(corpus2, content_transformer(tolower))
inspect(corpus2[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt  whipsaw wednesday  the view from dow  dow oil dollar
## [2] rt  whipsaw wednesday  the view from dow  dow oil dollar
## [3] rt  is making moves toward triple digit gains make your move today stockmarket paid ...
## [4] rt  whipsaw wednesday  the view from dow  dow oil dollar
## [5] rt  with behind us focus turns to tsla with hopes of more insight into the company's delivery sc
```

```
#Remove stop words
corpus2 <- tm_map(corpus2, removeWords, stopwords('english'))
inspect(corpus2[1:5])
```

```
#Remove White spaces
corpus2 <- tm_map(corpus2, stripWhitespace)
inspect(corpus2[1:5])
```

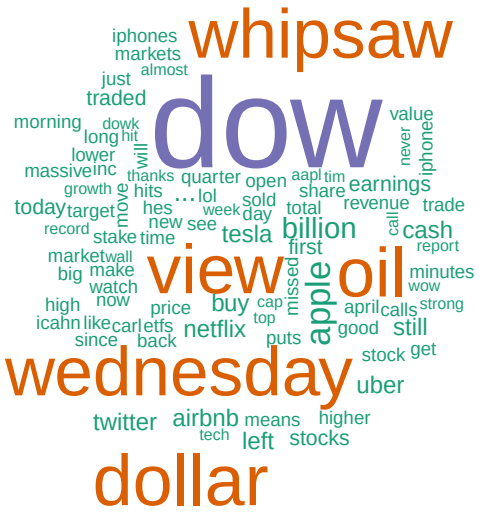
```
dtm2 <- DocumentTermMatrix(corpus2)
inspect(dtm2)
```

```
## <<DocumentTermMatrix (documents: 1000, terms: 1177)>>
## Non-/sparse entries: 6760/1170240
## Sparsity           : 99%
## Maximal term length: 19
## Weighting          : term frequency (tf)
## Sample             :
##      Terms
## Docs  ... apple billion buy dollar dow oil view wednesday whipsaw
##   313 0     0       0   0      0   0   0    0         0       0
##   328 0     0       0   0      0   0   0    0         0       0
##   332 1     1       0   0      0   0   0    0         0       0
##   350 0     1       0   0      0   0   0    0         0       0
##   361 0     0       0   0      0   0   0    0         0       0
##   406 0     1       0   0      0   0   0    0         0       0
##   418 1     0       0   0      0   0   0    0         0       0
##   423 0     0       0   0      0   0   0    0         0       0
##   567 1     0       0   0      0   0   0    0         0       0
##   944 0     0       0   0      0   0   0    0         0       0
```

```
dtm2 <- as.data.frame(as.matrix(dtm2))
freq2 = data.frame(sort(colSums(as.matrix(dtm2)), decreasing=TRUE))
#top 2 frequent words:
head(freq2,2)
```

```
##        sort.colSums.as.matrix.dtm2....decreasing...TRUE.
## dow                                                  918
## dollar                                               467
```

```
wordcloud(rownames(freq2), freq2[,1] , min.freq = 8, colors=brewer.pal(1, "Dark2"))
```



Q1.3

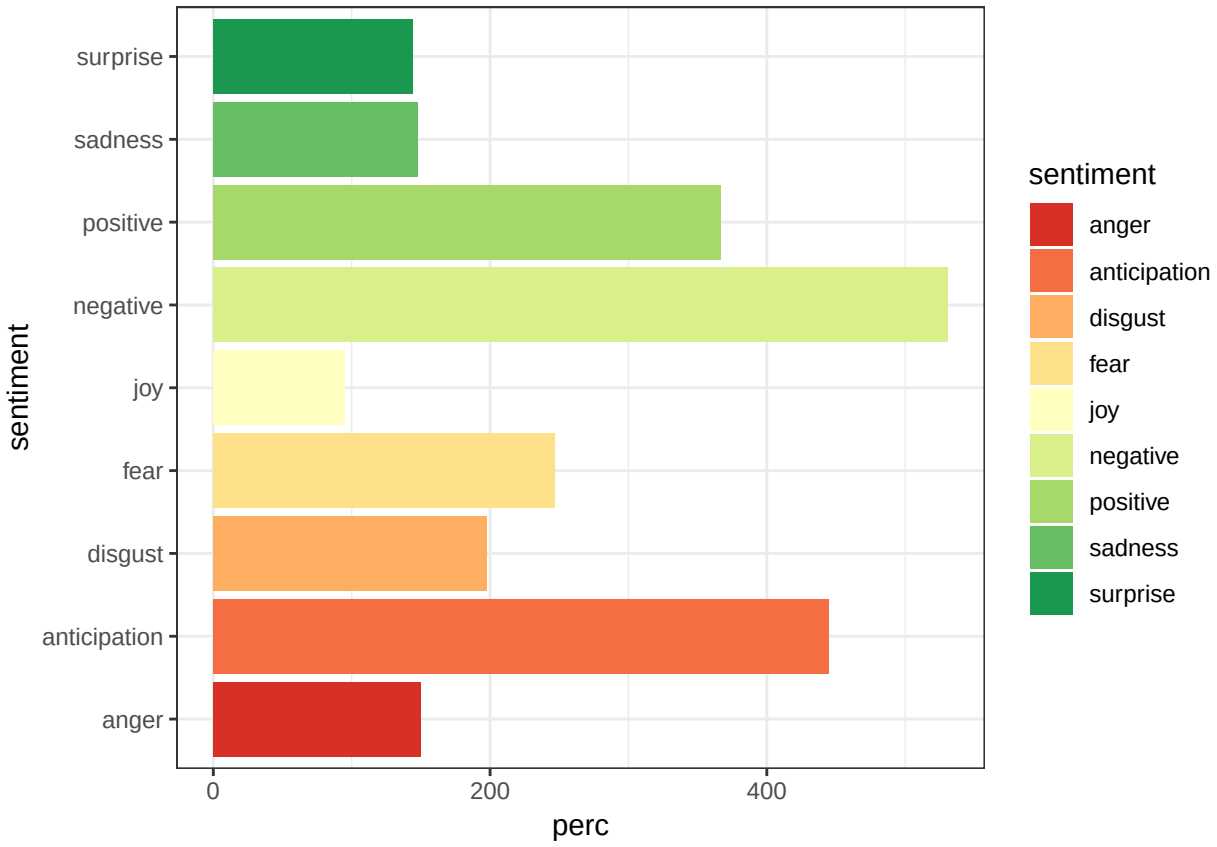
```
#let's calculate scores for our texts:
#Before Cleaning of the first dataset
scores <- get_nrc_sentiment(Apple1Text)
summary(scores)
```

```
##      anger       anticipation      disgust           fear
##  Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.00   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.15   Mean   :0.445   Mean   :0.198   Mean   :0.247
##  3rd Qu.:0.00   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :2.00   Max.   :3.000   Max.   :2.000   Max.   :2.000
##       joy           sadness         surprise         trust
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.095   Mean   :0.148   Mean   :0.144   Mean   :0.358
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.250
##  Max.   :3.000   Max.   :2.000   Max.   :2.000   Max.   :3.000
##     negative        positive
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000
##  Mean   :0.531   Mean   :0.367
##  3rd Qu.:1.000   3rd Qu.:1.000
```

```
##  Max.   :3.000   Max.   :3.000
```

```
scores$sentiment <- Apple1$sentiment
```

```
#Generate barplot
scores <- scores %>%
  summarise(
    anger = sum(anger),
    anticipation = sum(anticipation),
    disgust = sum(disgust),
    fear = sum(fear),
    joy = sum(joy),
    sadness = sum(sadness),
    surprise = sum(surprise),
    negative = sum(negative),
    positive = sum(positive))


scores_gathered <- scores %>%
  gather("sentiment", "value") %>%
  mutate(perc = value )

ggplot(scores_gathered, aes(x = sentiment, y = perc, fill = sentiment)) +
  geom_histogram(stat = "identity") +
  coord_flip() +
  theme_bw() +
  scale_fill_brewer(palette="RdYlGn")
```

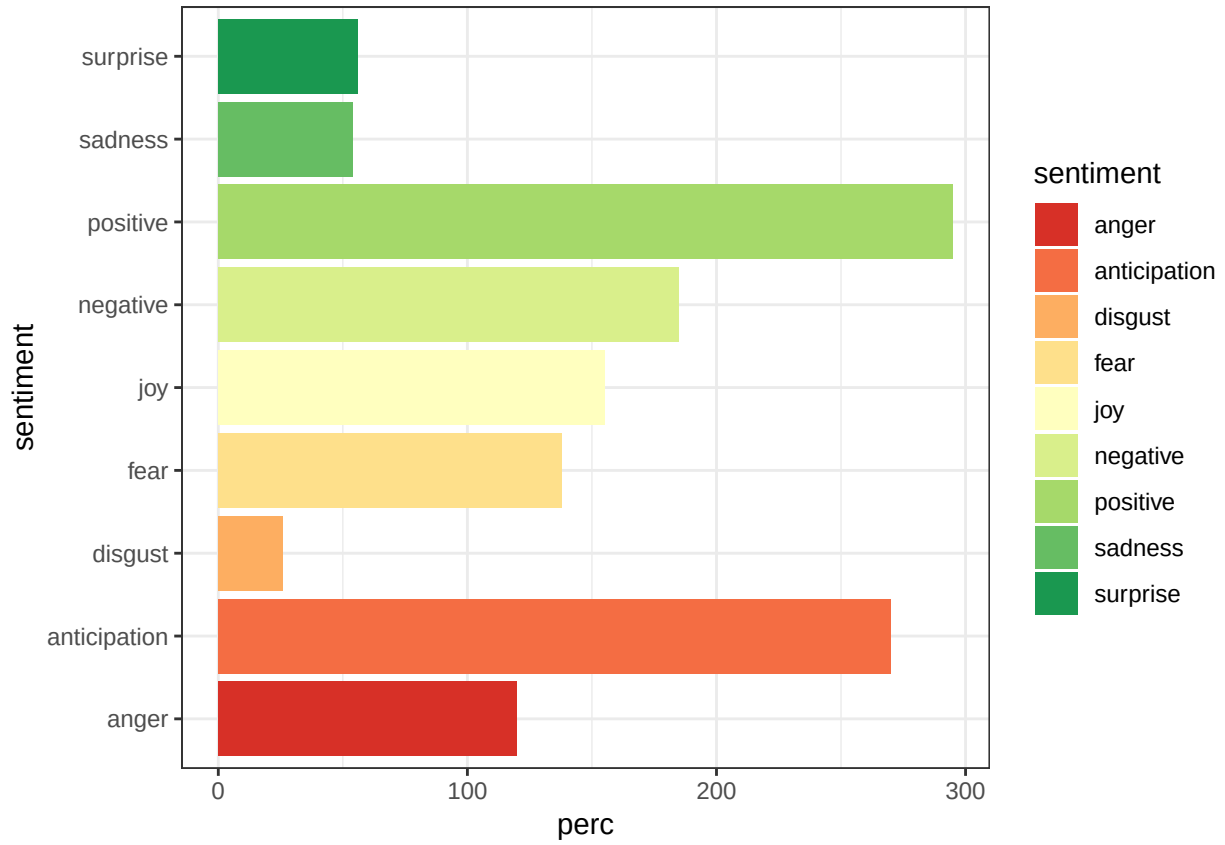


```r
#Before Cleaning of the second dataset
scores <- get_nrc_sentiment(Apple2Text)
summary(scores)
```

```
##      anger       anticipation     disgust           fear
##  Min.   :0.00   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.00   Median :0.00   Median :0.000   Median :0.000
##  Mean   :0.12   Mean   :0.27   Mean   :0.026   Mean   :0.138
##  3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :2.00   Max.   :3.00   Max.   :2.000   Max.   :3.000
##       joy           sadness         surprise         trust
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.155   Mean   :0.054   Mean   :0.056   Mean   :0.228
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :3.000   Max.   :1.000   Max.   :2.000   Max.   :4.000
##     negative        positive
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000
##  Mean   :0.185   Mean   :0.295
##  3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :3.000   Max.   :3.000
```

```r
scores$sentiment <- Apple2$sentiment
```

```r
scores <- scores %>%
  summarise(
    anger = sum(anger),
    anticipation = sum(anticipation),
    disgust = sum(disgust),
    fear = sum(fear),
    joy = sum(joy),
    sadness = sum(sadness),
    surprise = sum(surprise),
    negative = sum(negative),
    positive = sum(positive))


scores_gathered <- scores %>%
  gather("sentiment", "value") %>%
  mutate(perc = value )

ggplot(scores_gathered, aes(x = sentiment, y = perc, fill = sentiment)) +
  geom_histogram(stat = "identity") +
  coord_flip() +
  theme_bw() +
  scale_fill_brewer(palette="RdYlGn")
```



The positive value increased "after" the announcing the quarterly profits

Q1.4

```
#Combine datasets
apple3 <- rbind(Apple1, Apple2)
#Data cleaning and preparation for Random Forest Model
apple3 <- na.omit(apple3)
apple3Text <- iconv(apple3$text, to = "utf-8")
corpus3 <- Corpus(VectorSource(apple3Text))
corpus3 <- tm_map(corpus3, content_transformer(removeURL))
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [2] RT @option_snipper: $AAPL beat on both eps and revenues. SEES 4Q REV. $49B-$52B, EST. $49.1B
## [3] Let's see this break all timers. $AAPL 156.89
## [4] RT @SylvaCap: Things might get ugly for $aapl with the iphone delay. With $aapl down that means a
## [5] $AAPL - wow! This was supposed to be a throw-away quarter and AAPL beats by over 500 million in i
```

```
corpus3 <- tm_map(corpus3, content_transformer(removeat))
corpus3 <- tm_map(corpus3, content_transformer(removedollar))
#Remove Punctuation
corpus3 <- tm_map(corpus3, removePunctuation)
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  beat on both eps and revenues SEES 4Q REV  EST 1B
## [2] RT  beat on both eps and revenues SEES 4Q REV  EST 1B
## [3] Lets see this break all timers 15689
## [4] RT  Things might get ugly for with the iphone delay With down that means almost all of the FANG s
## [5]  wow This was supposed to be a throwaway quarter and AAPL beats by over 500 million in revenue Tr
```

```
#Remove Numbers
corpus3 <- tm_map(corpus3, removeNumbers)
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] RT  beat on both eps and revenues SEES Q REV  EST B
## [2] RT  beat on both eps and revenues SEES Q REV  EST B
## [3] Lets see this break all timers
## [4] RT  Things might get ugly for with the iphone delay With down that means almost all of the FANG s
## [5]  wow This was supposed to be a throwaway quarter and AAPL beats by over  million in revenue Trill
```

```
corpus3 <- tm_map(corpus3, content_transformer(tolower))
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt  beat on both eps and revenues sees q rev  est b
## [2] rt  beat on both eps and revenues sees q rev  est b
## [3] lets see this break all timers
## [4] rt  things might get ugly for with the iphone delay with down that means almost all of the fang s
## [5]  wow this was supposed to be a throwaway quarter and aapl beats by over  million in revenue trill
```

```
#Remove stop words
corpus3 <- tm_map(corpus3, removeWords, stopwords('english'))
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt  beat   eps  revenues sees q rev  est b
## [2] rt  beat   eps  revenues sees q rev  est b
## [3] lets see  break  timers
## [4] rt  things might get ugly    iphone delay    means almost    fang stocks   pos...
## [5]  wow   supposed    throwaway quarter  aapl beats    million  revenue trillion dollar company
```

```
#Remove Special Characters
specialChars<-function(x) gsub("[^[:alnum:][:blank:]?&/\\-]", "", x)
corpus3 <-tm_map(corpus3, specialChars)
#Remove White spaces
corpus3 <- tm_map(corpus3, stripWhitespace)
inspect(corpus3[1:5])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
## [1] rt beat eps revenues sees q rev est b
## [2] rt beat eps revenues sees q rev est b
## [3] lets see break timers
## [4] rt things might get ugly iphone delay means almost fang stocks pos
## [5]  wow supposed throwaway quarter aapl beats million revenue trillion dollar company
```

```
#Document term matrix of the combined dataset
dtm3 <- DocumentTermMatrix(corpus3)
cleanedcorpus3 <- as.data.frame(as.matrix(dtm3))
colnames(cleanedcorpus3) <- make.names(colnames(cleanedcorpus3))
cleanedcorpus3$label <- apple3$sentiment
#Split the dataset to 80-20, train and test set.
train_idx <- sample(nrow(cleanedcorpus3), round(nrow(cleanedcorpus3)/100*80,0), replace = F)
train <- cleanedcorpus3[train_idx,]
test <- cleanedcorpus3[-train_idx,]
```

```
train <- na.omit(train)
#Convert the labels column to factor
train$label <- as.factor(train$label)
#train the random forest model
rf <- randomForest(label~., train)
#Predict using the test set
prediction <- predict(rf, test)
#build the Confusion Matrix
confMatrix <- as.matrix(table(test$label,prediction))
 n = sum(confMatrix) # number of instances
 nc = nrow(confMatrix) # number of classes
 diag = diag(confMatrix) # number of correctly classified instances per class
 rowsums = apply(confMatrix, 1, sum) # number of instances per class
 colsums = apply(confMatrix, 2, sum) # number of predictions per class
 p = rowsums / n # distribution of instances over the actual classes
 q = colsums / n # distribution of instances over the predicted classes
# Accuracy is the diagonal summation over the total count
accuracy = sum(diag) / n
#Model Accuracy:
accuracy
```

```
## [1] 0.8375
```

```
precision = diag / colsums
# Model Precesion
precision
```

```
##  negative   neutral  positive
```

```
## 0.9315068 0.8000000 0.9148936
```

```
recall = diag / rowsums
#Model Recall
recall
```

```
##  negative   neutral  positive
## 0.7391304 0.9824561 0.5375000
```

```
f1 = 2 * precision * recall / (precision + recall)
data.frame(precision, recall, f1)
```

```
##          precision    recall        f1
## negative 0.9315068 0.7391304 0.8242424
## neutral  0.8000000 0.9824561 0.8818898
## positive 0.9148936 0.5375000 0.6771654
```

```
#one-vs-all confusion matrix for each class
oneVsAll = lapply(1 : nc,
                  function(i){
                    v = c(confMatrix[i,i],
                          rowsums[i] - confMatrix[i,i],
                          colsums[i] - confMatrix[i,i],
                          n-rowsums[i] - colsums[i] + confMatrix[i,i]);
                    return(matrix(v, nrow = 2, byrow = T))})
oneVsAll
```

```
## [[1]]
##      [,1] [,2]
## [1,]   68   24
## [2,]    5  303
##
## [[2]]
##      [,1] [,2]
## [1,]  224    4
## [2,]   56  116
##
## [[3]]
##      [,1] [,2]
## [1,]   43   37
## [2,]    4  316
```

```
s = matrix(0, nrow = 2, ncol = 2)
for(i in 1 : nc){s = s + oneVsAll[[i]]}
#Summing up the values of these 3 matrices results in one confusion matrix
s
```

```
##      [,1] [,2]
## [1,]  335   65
## [2,]   65  735
```

Part2.

```
library("igraph")
library("ggplot2")
```

```
# read the data

links <- read_csv("/home/nesma/SemesterII/BusinessDataAnalytics/HW6/Hi-tech-Edges.csv")
View(links)
str(links)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 129 obs. of  3 variables:
##  $ from  : num  10 28 2 2 2 23 23 15 15 15 ...
##  $ to    : num  2 2 10 4 29 24 29 29 14 34 ...
##  $ weight: num  10 24 36 47 4 28 38 17 8 16 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   from = col_double(),
##   ..   to = col_double(),
##   ..   weight = col_double()
##   .. )
```

```
nodes <- read_csv("/home/nesma/SemesterII/BusinessDataAnalytics/HW6/Hi-tech-Nodes.csv")
View(nodes)
str(nodes)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 36 obs. of  4 variables:
##  $ Node      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name      : chr  "Abe" "Bob" "Carl" "Dale" ...
##  $ Gender    : chr  "male" "male" "male" "male" ...
##  $ Department: chr  "Management" "Marketing" "Development" "Management" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Node = col_double(),
##   ..   Name = col_character(),
##   ..   Gender = col_character(),
##   ..   Department = col_character()
##   .. )
```

```
cat("Amount of rows in nodes data: ", nrow(nodes), "\n")
```

```
## Amount of rows in nodes data:  36
```

```
cat("Amount of unique nodes: ", length(unique(nodes$Node)), "\n")
```

```
## Amount of unique nodes:  36
```

```
cat("Amount of rows in links data: ", nrow(links), "\n")
```

```
## Amount of rows in links data:  129
```

```
cat("Amount of unique links: ", nrow(unique(links[,c("from", "to")])), "\n")
```

```
## Amount of unique links:  129
```

Build a directied graph:

```
net <- graph_from_data_frame(d=links, vertices=nodes, directed=T)
net
```

```
## IGRAPH a329cad DNW- 36 129 --
## + attr: name (v/c), Name (v/c), Gender (v/c), Department (v/c),
## | weight (e/n)
## + edges from a329cad (vertex names):
```

```
##  [1] 10->2  28->2  2 ->10 2 ->4  2 ->29 23->24 23->29 15->29 15->14 15->34
## [11] 7 ->4  7 ->24 14->2  14->7  14->15 34->15 34->14 34->29 34->24 34->20
## [21] 29->23 29->7  29->2  29->18 29->11 29->20 29->9  29->34 29->14 29->15
## [31] 18->27 18->13 18->29 27->18 27->4  27->24 4 ->2  4 ->27 4 ->13 4 ->35
## [41] 4 ->24 13->18 13->16 13->30 13->29 13->4  13->2  24->4  24->30 24->21
## [51] 24->20 24->11 24->29 24->7  11->18 11->33 11->20 11->34 11->14 20->29
## [61] 20->11 20->4  20->24 20->13 20->33 20->21 20->26 20->22 20->34 22->34
## + ... omitted several edges
```

```
E(net) # The edges of the "net" object
```

```
## + 129/129 edges from a329cad (vertex names):
##   [1] 10->2  28->2  2 ->10 2 ->4  2 ->29 23->24 23->29 15->29 15->14 15->34
##  [11] 7 ->4  7 ->24 14->2  14->7  14->15 34->15 34->14 34->29 34->24 34->20
##  [21] 29->23 29->7  29->2  29->18 29->11 29->20 29->9  29->34 29->14 29->15
##  [31] 18->27 18->13 18->29 27->18 27->4  27->24 4 ->2  4 ->27 4 ->13 4 ->35
##  [41] 4 ->24 13->18 13->16 13->30 13->29 13->4  13->2  24->4  24->30 24->21
##  [51] 24->20 24->11 24->29 24->7  11->18 11->33 11->20 11->34 11->14 20->29
##  [61] 20->11 20->4  20->24 20->13 20->33 20->21 20->26 20->22 20->34 22->34
##  [71] 22->11 22->20 9 ->29 9 ->20 21->9  21->20 29->21 21->19 21->6  33->24
##  [81] 33->35 33->20 33->34 33->14 33->11 35->33 35->4  35->30 35->16 35->19
##  [91] 35->12 35->26 30->13 30->19 30->35 30->11 30->24 16->36 16->19 16->35
## + ... omitted several edges
```

```
V(net) # The vertices of the "net" object
```

```
## + 36/36 vertices, named, from a329cad:
##  [1] 1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23
## [24] 24 25 26 27 28 29 30 31 32 33 34 35 36
```

Q 2.1

```
# a. Density of the network;
edge_density(net)
```

```
## [1] 0.102381
```

```
# b. Clustering coefficient;
transitivity(net, type="global")
```

```
## [1] 0.3545455
```

```
# c. Reciprocity of the network;
reciprocity(net)
```

```
## [1] 0.6666667
```

```
# d. Average path length;
mean_distance(net, directed=T)
```

```
## [1] 2.638542
```

```
# e. Diameter (by considering weights).
diameter(net, directed=T, weights=E(net)$weight)
```

```
## [1] 102
```

1. From the reciprocity value we can colnclude that more than half the vertices are mutually linked to each other.

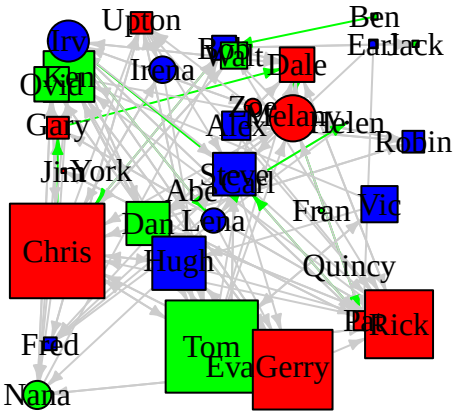
2. From the edge density, it reurns small number which means that the possibility of a fully connected graph is small.
3. From the transitivity (Clustering Coefficient), it returns a small number which means that small number of adjacent vertices are connected.
4. The average path length between each pair of nodes in the graph is 2.63 which is a resinable distance.

Q 2.2

```
#helping resources: http://www.shizukalab.com/toolkits/sna/plotting-networks-pt-2
#https://kateto.net/netscix2016.html
#Each node has different color, depending on department to which user belongs;
V(net)$color=V(net)$Department
V(net)$color=gsub("Management","red",V(net)$color) #Management will be red
V(net)$color=gsub("Marketing","blue",V(net)$color) #Marketing will be blue
V(net)$color=gsub("Development","green",V(net)$color) #Development will be green


hs <- hub_score(net, weights=NA)$vector  #- Each node has different size depending on the hub size of t

diam <- get_diameter(net, directed=T)
diam
```

```
## + 12/36 vertices, named, from a329cad:
##  [1] 28 2  29 7  4  13 16 19 35 12 3  8
```

```
#- Find the path of the diameter on the graph and colorize its edges only.
ecol <- rep("gray80", ecount(net))
ecol[E(net, path=diam)] <- "green"

plot(net,edge.arrow.size=.4, remove.multiple = T, remove.loops = T,layout=layout_randomly,
     edge.size= E(net)$weight,        #Each edge has different size depending on the weight;
     vertex.shape = ifelse(V(net)$Gender == "female", "circle", "square"),  #- Each node has different
     vertex.size=hs*50, #- Each node has different size depending on the hub size of the nodes;
     vertex.label=V(net)$Name,        #Each node has a name associated with it;
     vertex.label.color="black",
     edge.color = ecol)
```



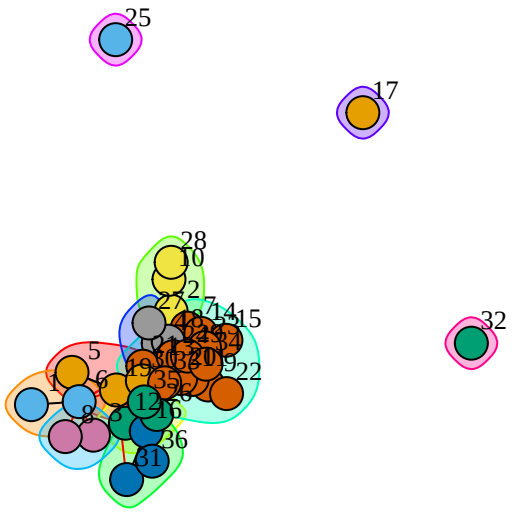
Q 2.3

```
wtc <- walktrap.community(as.undirected(net),modularity = T)
wtc
```

```
## IGRAPH clustering walktrap, groups: 11, mod: 0.35
## + groups:
##   $`1`
##   [1] "5"  "19" "30"
##
##   $`2`
##   [1] "1" "6"
##
##   $`3`
##   [1] "12" "26" "35"
##
##   $`4`
##   + ... omitted several groups/vertices
```

```
set.seed(50)
plot(wtc, net,
     edge.arrow.mode=0,
     edge.arrow.size=.2,
     vertex.label.dist=2,
     # vertex.label=V(net)$Name,
     vertex.label.cex=.8,
     vertex.label.color="black"
     # ,layout=layout_components
```

)



1. The second cluster (light Blue Cluster) contains “1”, “6”
2. Checking the Links dataset we have from 1 to 6 the weight is 16 and it’s the only node that 1 is connected to it for this sample it make sense.
3. For the third cluster 12 and 26 appear, checking the links table, we found that from 12 to 26 has weight of 25 which is the highest weight from 12