Week7_lab

Frank Shen

2019/3/4

```
string = c("bach", "back", "beech", "beach","black")
grep("be(a|e)ch", string, value = TRUE)
```

```
## [1] "beech" "beach"
```

```
grep("be[ae]ch", string, value = TRUE)
```

```
## [1] "beech" "beach"
```

```
grep("b(e+|a+)ch", string, value = TRUE)
```

```
## [1] "bach"  "beech"
```

```
grep("b[ae]e?ch", string, value = TRUE)
```

```
## [1] "bach"  "beech"
```

```
library(twitteR)
library(dplyr)
```

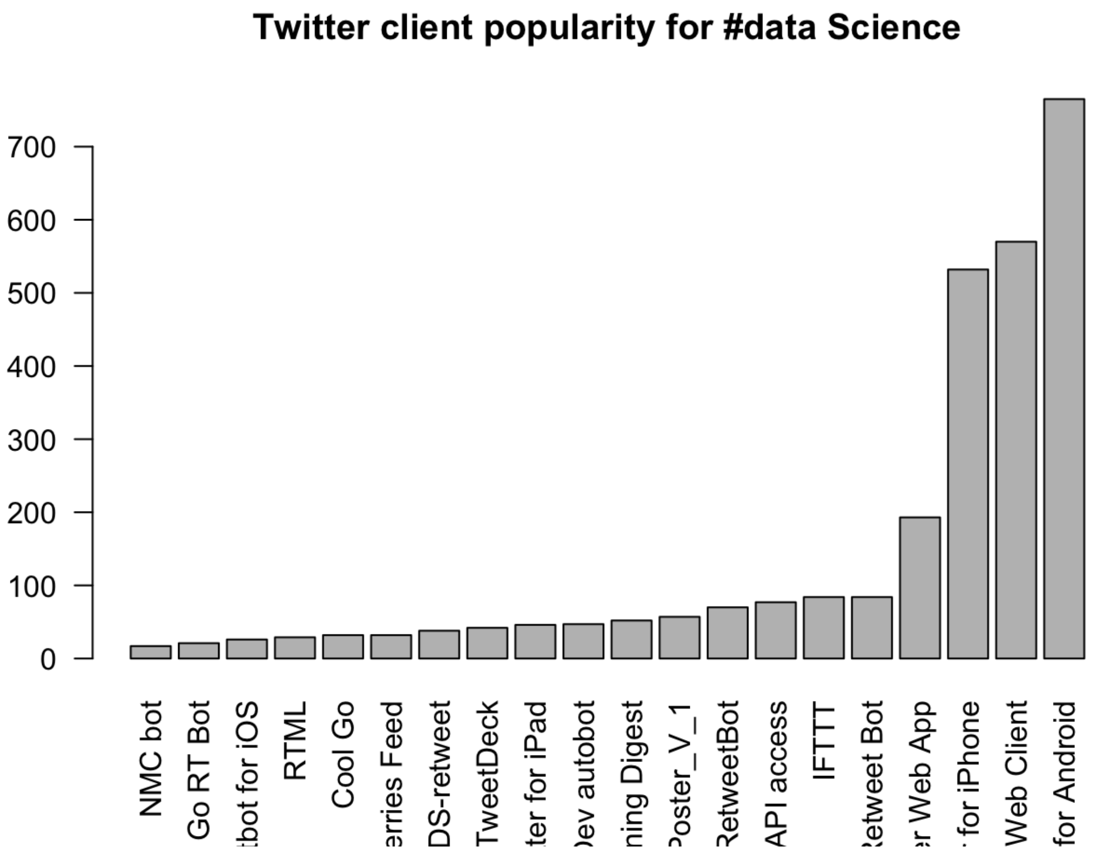
```
##
## Attaching package: 'dplyr'
```

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

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

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

```
##Question1
dsTweets = load("Tweets_Lab7.Rdata")
dsTweetsDF = twListToDF(Tweets2Use)
dsTweetsDF = dsTweetsDF[1:3200 ,]
temp = gsub("</a>", "", dsTweetsDF$statusSource, perl = TRUE)
platform = gsub("<.*>", "", temp, perl = TRUE)
barplot(tail(sort(table(platform)), 20),las = 2, main = "Twitter client popularity for #data
Science")
```



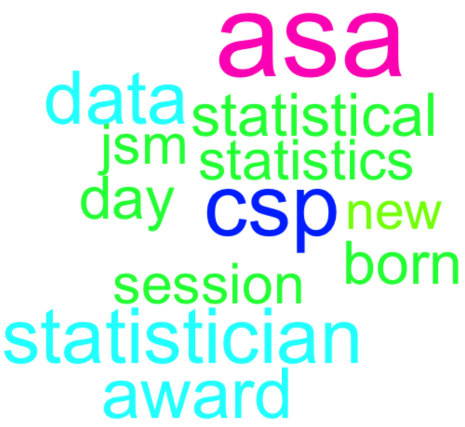
```
##Question 2
load("NewUserLab7.Rdata")
Userlab = twListToDF(UserTweets)
Userlab = Userlab[1:500 ,]
sentence = Userlab$text
sentence = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", sentence)
sentence = gsub("@\\w+", " ", sentence)
sentence = gsub("(?!')[[:punct:]]", "", sentence, perl = T)
sentence = gsub("[[:cntrl:]]", "", sentence)
sentence = gsub("[[:digit:]]", "", sentence)
sentence = iconv(sentence, "ASCII", "UTF-8", sub = "")
sentence = tolower(sentence)
sentence = gsub("http\\w+", "", sentence)
sentence = gsub("[ \t]{2,}", " ", sentence)
sentence = gsub("^\\s+|\\s+$", "", sentence)
word.list = strsplit(sentence, " ")
words = unlist(word.list)
library(tm)
```

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

```
words = words[!words %in% tm::stopwords(kind = "english")]
library(wordcloud)
```

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

```
wordcloud(names(table(words)), table(words), colors = rainbow(8), min.freq = 10)
```



```
library(tm)
pos = scan("positive-words.txt", what = "character", comment.char = ";")
neg = scan("negative-words.txt", what = "character", comment.char = ";")

length(pos)
```

```
## [1] 2006
```

```
length(neg)
```

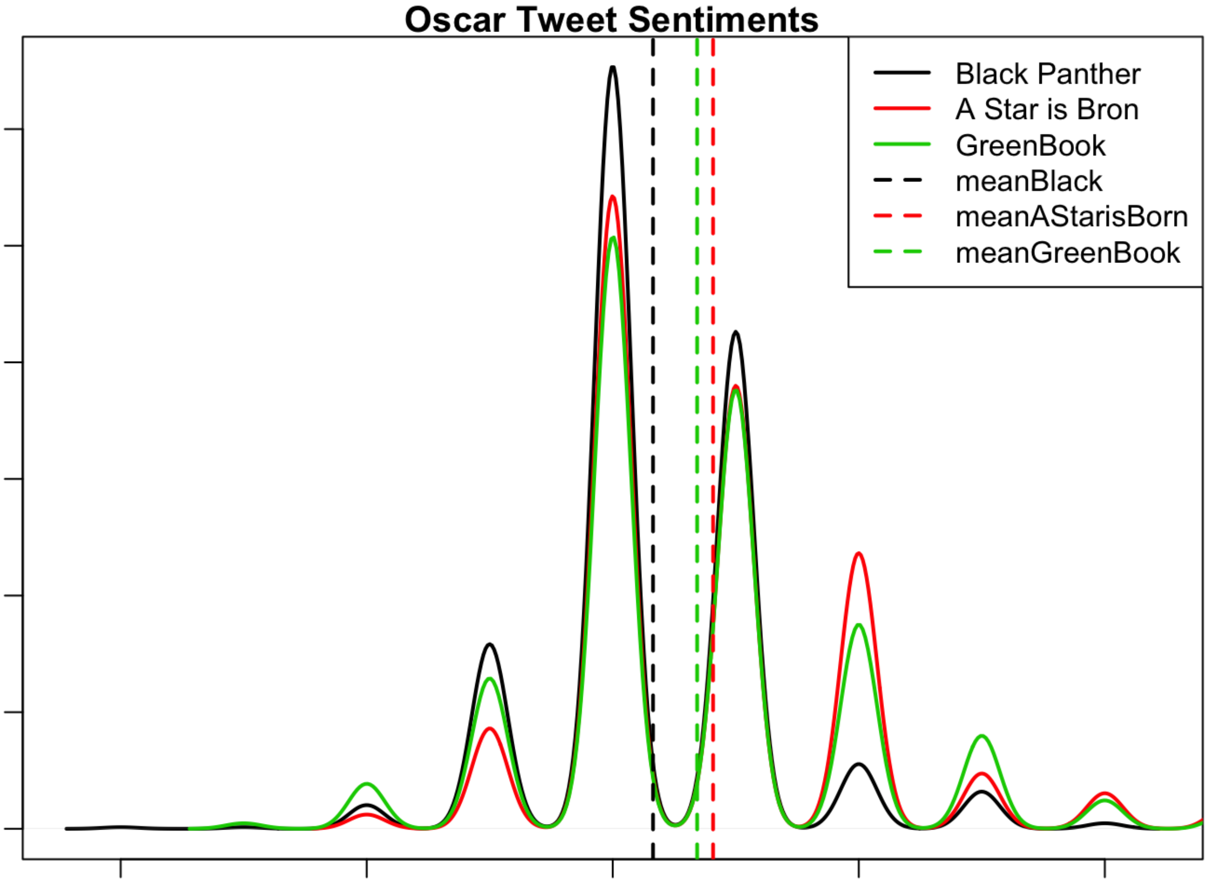
```
## [1] 4783
```

```
neg = c(neg, "wtf")
getSentimentScore = function(tweet_text, pos, neg) {
sentence = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweet_text)
sentence = gsub("@\\w+", "", sentence)
sentence = gsub("[[:punct:]]", "", sentence)
sentence = gsub("[[:cntrl:]]", "", sentence)
sentence = gsub("[[:digit:]]", "", sentence)
sentence = gsub("http\\w+", "", sentence)
sentence = gsub("^\\s+|\\s+$", "", sentence)
sentence = iconv(sentence, "ASCII", "UTF-8", sub = "")
sentence = tolower(sentence)
word.list = strsplit(sentence, " ")
score = numeric(length(word.list))
for (i in 1:length(word.list)) {
  pos.matches = match(word.list[[i]], pos)
  neg.matches = match(word.list[[i]], neg)
  pos.matches = !is.na(pos.matches)
  neg.matches = !is.na(neg.matches)
  score[i] = sum(pos.matches) - sum(neg.matches)
  }
return(score)
}


##Question3 a)
load("SentimentTweets.Rdata")

BlackDF = twListToDF(T1)
BlackDF = BlackDF[1:2000,]
StarDF = twListToDF(T2)
StarDF = StarDF[1:2000,]
GreenDF = twListToDF(T3)
GreenDF = GreenDF[1:2000,]
##Question3 b)
BlackSent = getSentimentScore(BlackDF$text, pos, neg)
StarSent = getSentimentScore(StarDF$text, pos, neg)
GreenSent = getSentimentScore(GreenDF$text, pos, neg)

par(mfrow = c(1, 1))
par(mar = c(1, 1, 1, 1))
plot(density(BlackSent), col = 1, main = "Oscar Tweet Sentiments", xlab = "Sentiment Score",
lwd = 2)
lines(density(StarSent), col = 2, lwd = 2)
lines(density(GreenSent), col = 3, lwd = 2)
abline(v = mean(BlackSent), col = 1, lwd = 2, lty = 2)
abline(v = mean(StarSent), col = 2, lwd = 2, lty = 2)
abline(v = mean(GreenSent), col = 3, lwd = 2, lty = 2)
legend("topright", c("Black Panther", "A Star is Bron", "GreenBook", "meanBlack", "meanAStar
isBorn", "meanGreenBook"), col = c(1:3, 1:3),lwd = 2, lty = c(1, 1, 1, 2, 2, 2))
```



```
## Question3 C)
sum(BlackSent > 0)/length(BlackSent)
```

```
## [1] 0.383
```

```
sum(StarSent > 0)/length(StarSent)
```

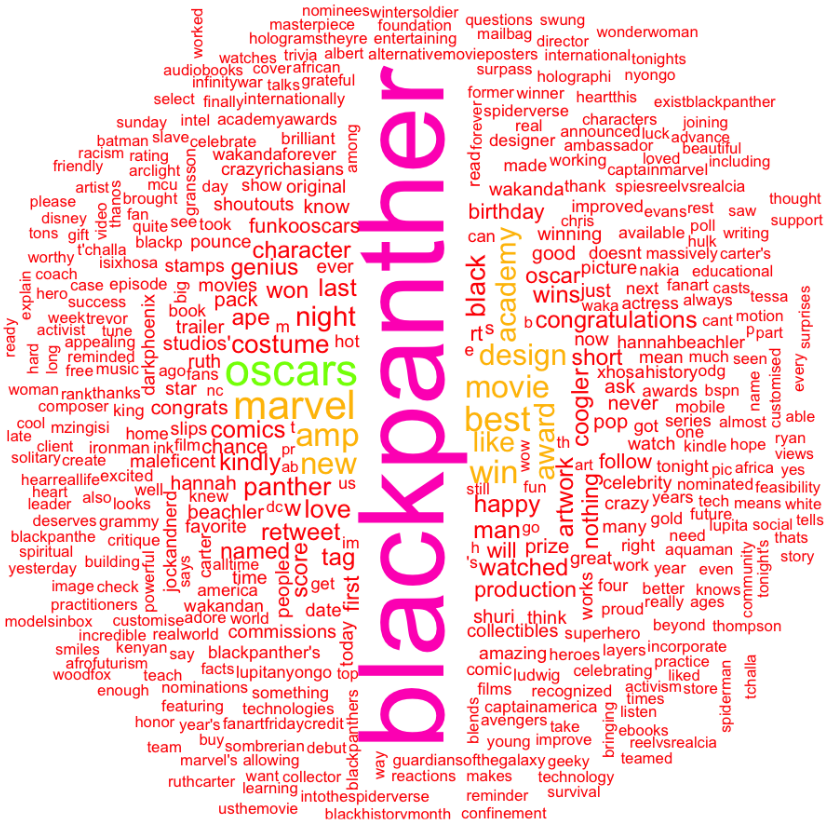
```
## [1] 0.527
```

```
sum(GreenSent > 0)/length(GreenSent)
```

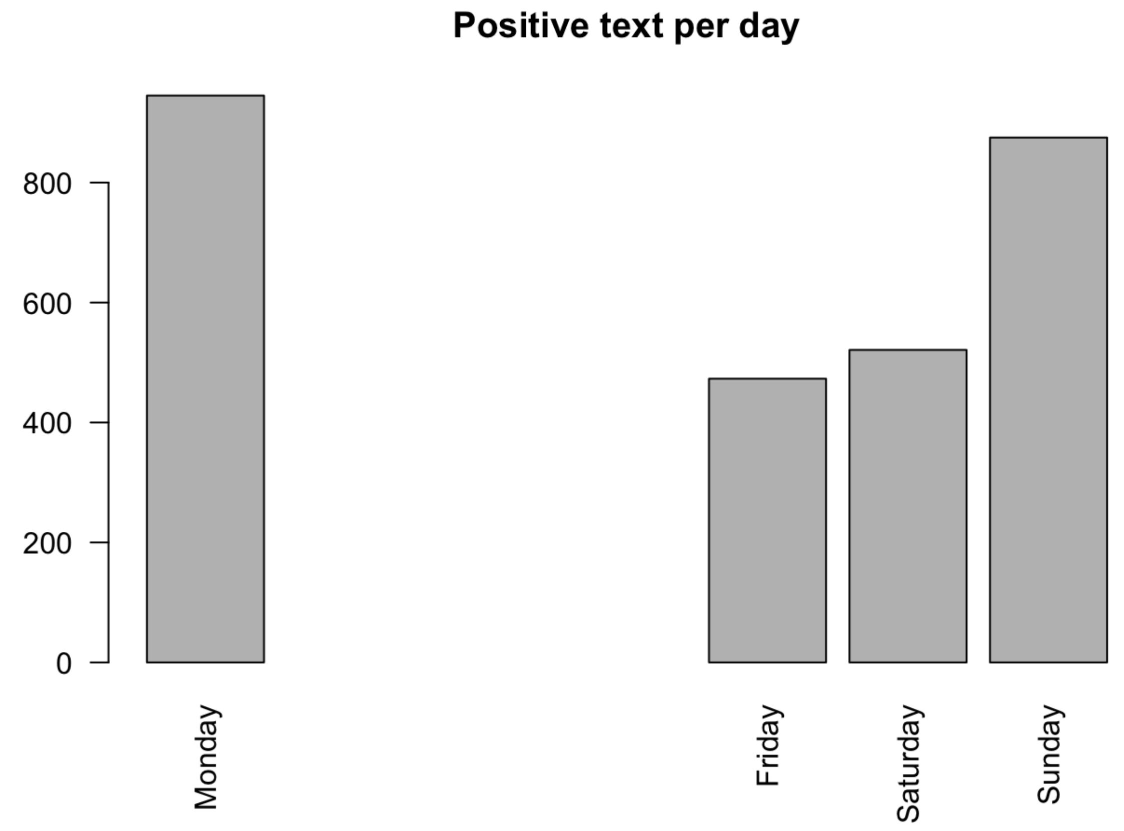
```
## [1] 0.497
```

```
##Star has highest propotion

PosStarText = StarDF$text[StarSent > 0]
sentence2 = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", PosStarText)
sentence3 = gsub("@\\w+", " ", sentence2)
sentence4 = gsub("(?!')[[:punct:]]", "", sentence3, perl = T)
sentence5 = gsub("[[:cntrl:]]", "", sentence4)
sentence6 = gsub("[[:digit:]]", "", sentence5)
sentence7 = iconv(sentence6, "ASCII", "UTF-8", sub = "")
sentence8 = tolower(sentence7)
sentence9 = gsub("http\\w+", "", sentence8)
sentence10 = gsub("[ \t]{2,}", " ", sentence9)
sentence11 = gsub("^\\s+|\\s+$", "", sentence10)
word.list = strsplit(sentence11, " ")
words = unlist(word.list)
words = words[!words %in% tm::stopwords(kind = "english")]
freq = table(words)
library(wordcloud)
wordcloud(names(freq), freq, min.freq = 5, colors = rainbow(8), random.order = FALSE)
```



```
##Question3 D)
par(mar = c(5,4,4,2))
textAll = rbind(BlackDF, StarDF, GreenDF)
sentAll = c(BlackSent, StarSent, GreenSent)
postext = textAll[sentAll > 0 , ]
barplot(table(weekdays(postext[,"created"]))[c("Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday", "Sunday")], las = 2, main = "Positive text per day")
```



```
##This data might misleading us because we didn't get the days from Tuesdays to Thursdays. T
hus, we can not know that which day has the most of says about these movies
```