Homework: Text Mining

This homework sheet will test your knowledge of text mining in R. 15

0

a) Load the package `tm` and create a corpus from the twitter tweets that were sent on Election Day 2012.

```
library(tm)
```

```
elections <- as.data.frame(read.csv("elections2012.csv", header = TRUE, sep = ",",
    stringsAsFactors = FALSE))
set.before <- substring(elections$entry.CreatedAt, 1, 15) == "11/07/2012 04:5"
set.after <- substring(elections$entry.CreatedAt, 1, 15) == "11/07/2012 05:0"
elections <- elections[set.before | set.after, ]
elections <- elections[order(elections$entry.CreatedAt), ]
elections.corpus <- Corpus(VectorSource(elections$entry.Text))
```

1

b) The final poll closing times on Election Day were 11 p.m. EST at the west coast (except Hawaii). Convert this time into CET (Central European Time).

Solution:
By looking at the world map, you see that there is a time difference of 6 hours. Thus, the correct answer is November 7, 5 a. m. Alternatively, you can use R to do the time conversion.

```
date <- ISOdate(2012, 11, 6, 23, 0, 0, tz = "EST")
date.eu <- as.POSIXlt(date, format = "%Y/%m/%d %H:%M", tz = "CET")
date.eu

## [1] "2012-11-07 05:00:00 CET"
```

The above code extracts the tweets shortly before and after poll closing.

1

c) Transform the corpus by stripping the whitespaces, removing numbers and punctuation and setting all letters to lower cases.

Solution:

```
elections.corpus <- tm_map(elections.corpus, stripWhitespace)
elections.corpus <- tm_map(elections.corpus, tolower)
elections.corpus <- tm_map(elections.corpus, removeNumbers)
elections.corpus <- tm_map(elections.corpus, removePunctuation)
```

1

d) Remove English stopwords and transform your corpus to a plain text.

Solution:

```
elections.corpus <- tm_map(elections.corpus, removeWords, stopwords("english"))
if (packageVersion("tm")$minor <= 5) {
    elections.corpus <- tm_map(elections.corpus, PlainTextDocument)
} else {
    elections.corpus <- tm_map(elections.corpus, PlainTextDocument)
}
```

1

e) Perform stemming on your corpus.

Solution:

```
elections.corpus <- tm_map(elections.corpus, stemDocument, language = "english")
```

1

f) Calculate the term-document (TDM) matrix corresponding to your corpus.

Solution:

```
tdm <- TermDocumentMatrix(elections.corpus)

## Warning: invalid document identifiers
```

1

g) Give the words, that occur at least 100 times in your TDM.

Solution:

```
findFreqTerms(tdm, 100)

## [1] "elect"  "obama"  "romney"
```

1

h) Remove the terms, that occur in less than 40 % of your documents.

Solution:

```
tdm.small <- removeSparseTerms(tdm, 0.4)
```

1

i) Use k-means to cluster the data with $k = 2$. Show a few exemplary tweets of each cluster.

Solution:

```
k <- kmeans(t(tdm.small), 2)
```

```
head(elections[k$cluster == 1, "entry.Text"])

## [1] "Great race. #NPR #Elections2012 http://t.co/UWDyiqhj"
## [2] "#TeamObama we did it! :) Oh what a great night! :) #Elections2012 now we need that #Immigration ref
## [3] "RT @AGayReality: Obama helps millions, Romney helps millionaires. #Elections2012"
## [4] "#Elections2012 #Obama 281 votes, #MittRomney 203.. That's what I call a Punch in da' face! #4MoreYe
## [5] "#Obama  #Romany  #Elections2012   ?????????? ???????????? ?????? ?????? ???????????? : ???????????
## [6] "RT @N24_de: +EIL+ NBC: Barack Obama gewinnt die US-Pr??sidentschaftswahl. Mehr auf http://t.co/IPnJ

head(elections[k$cluster == 2, "entry.Text"])

## [1] "I #HOPE #Moodys doesn't downgrade us based on an election, #businessesknowbest #elections2012 @TheJ
## [2] "The People of The United States are the big winners of the Presidential election! Congratulations!!
## [3] "RT @blogdiva: UNFUCKINGBELIEVABLE that @KarlRove is literally saying the elections hinge on CAYUGA
## [4] "Why is everyone interested in the USA #elections2012 when they can't be bothered with the UK electi
## [5] "2012 Election Post #4 http://t.co/rY91v1O0 #elections2012"
## [6] "RT @smwaqas: What If we Get Our President elected again for more 5 years? =O Even the thought is fr
```

1

j) Give an interpretation of the size of the cluster and the within-cluster sum of squares.

Solution:

```
k$size

## [1] 1077   76

sum(k$tot.withinss)

## [1] 105.5
```

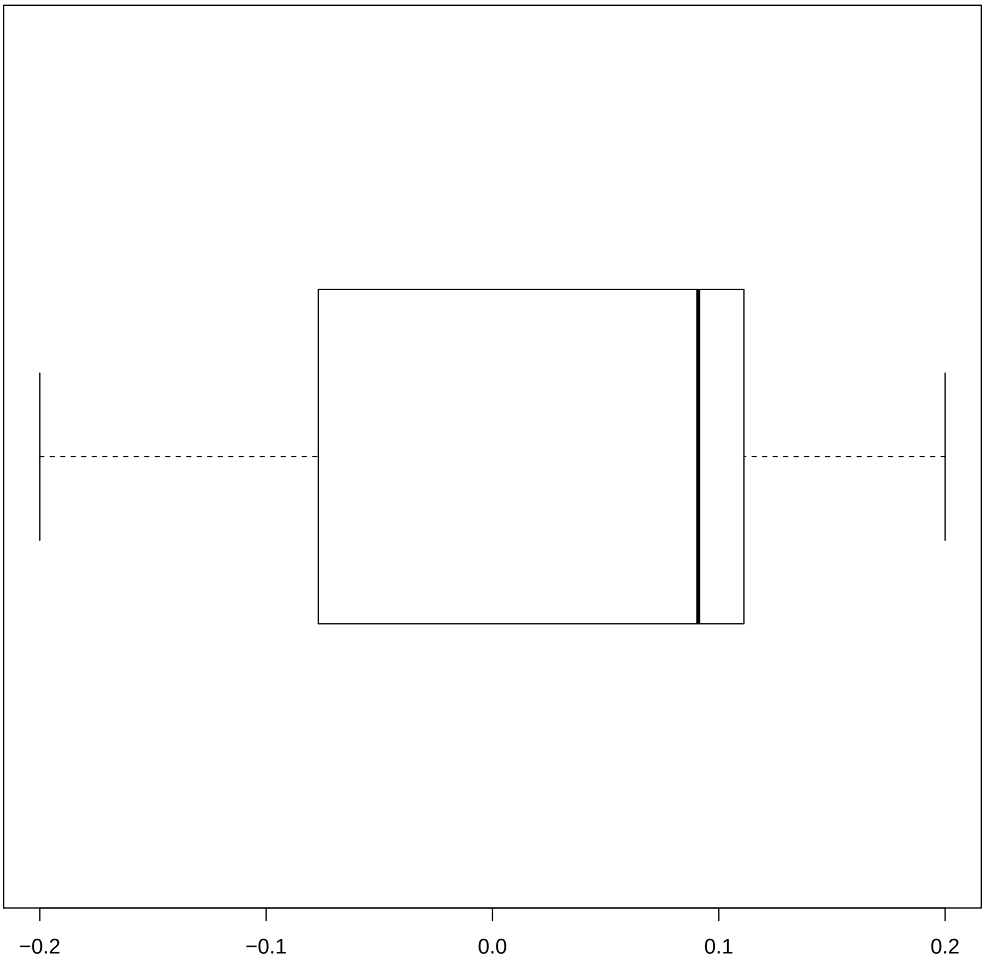
There is a larger and a smaller cluster. Maybe one cluster matches tweets addressing Obama and one addressing Romney (and elections in general). As Obama was presumably voted by younger people, a larger portion of tweets is devoted to him.

3

k) Calculate the Net-Optimism sentiment of the corpus and plot all non-zero sentiment values as a box-plot. You need to create a TDM of the positive and negative words to do so.

Solution:

```
pos <- as.data.frame(read.csv("positivity.txt", header = FALSE))
tdm.pos <- TermDocumentMatrix(elections.corpus, list(dictionary = t(pos)))

## Warning: invalid document identifiers

neg <- as.data.frame(read.csv("negativity.txt", header = FALSE))
tdm.neg <- TermDocumentMatrix(elections.corpus, list(dictionary = t(neg)))

## Warning: invalid document identifiers

N <- length(elections.corpus)
sentiment <- numeric(N)
for (i in 1:N) {
    sentiment[i] <- (sum(tdm.pos[, i]) - sum(tdm.neg[, i]))/sum(tdm[, i])
}

boxplot(sentiment[sentiment != 0], horizontal = TRUE)
```



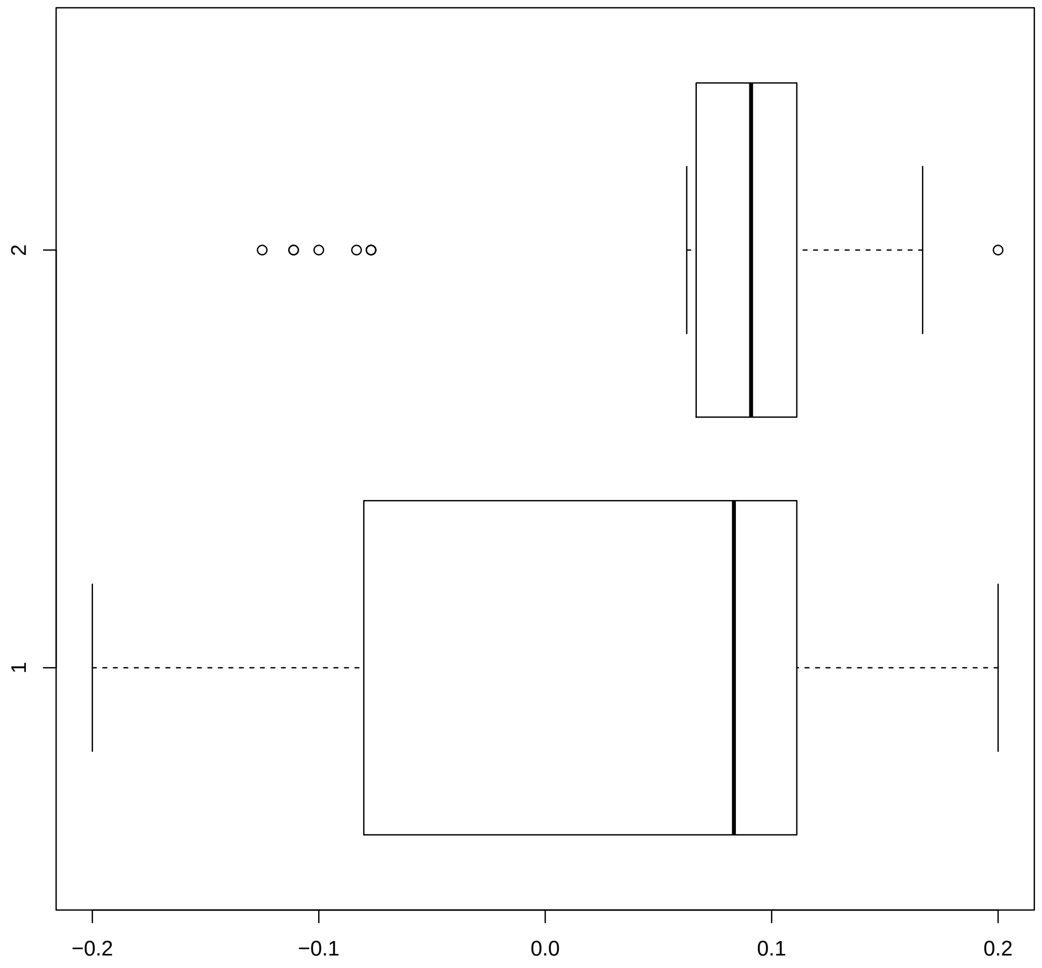
There are more documents with a positive Net-Optimism sentiment than with a negative one.

3

l) Is there a difference in the non-zero sentiment score before and after poll closing. Analyze visually!

Solution:

```
sentiment.before <- sentiment[substring(elections$entry.CreatedAt, 1, 15) == "11/07/2012 04:5"]
sentiment.after <- sentiment[substring(elections$entry.CreatedAt, 1, 15) == "11/07/2012 05:0"]

boxplot(sentiment.before[sentiment.before != 0],
        sentiment.after[sentiment.after != 0],
        horizontal=TRUE)
```



or

```
sentiment.before <- sentiment[1:622]
sentiment.after <- sentiment[623:1153]

boxplot(sentiment.before[sentiment.before != 0], sentiment.after[sentiment.after !=
    0], horizontal = TRUE)
```

There are more documents with a positive Net-Optimism sentiment than with a negative one.