Valiveti Manikanta Bhuvanesh

19BCD7088

Experiment 7

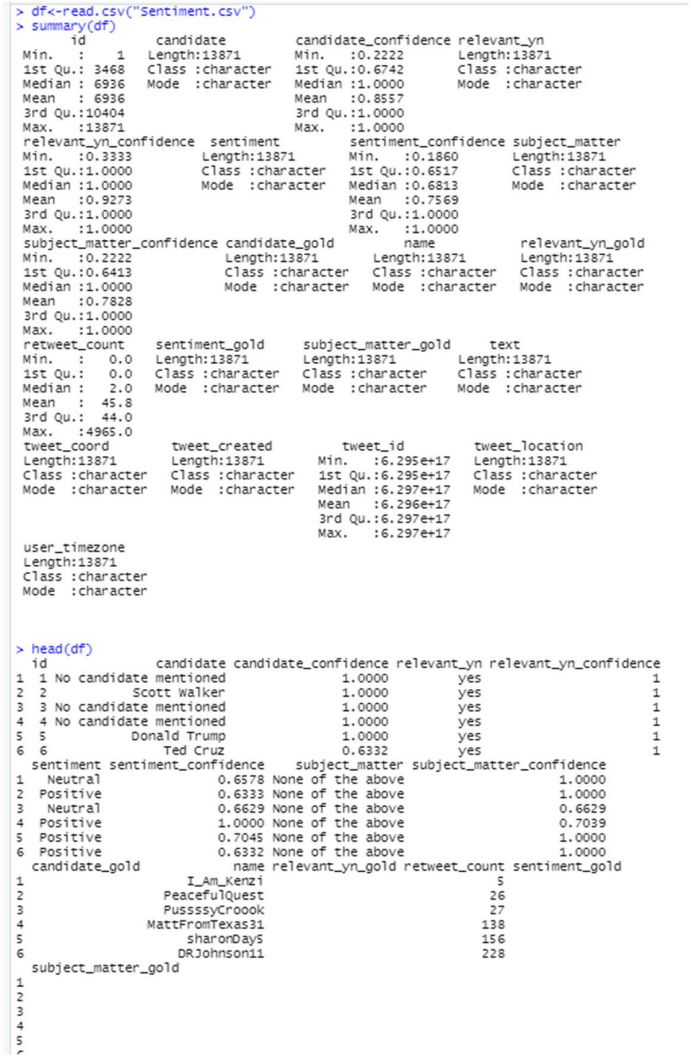
AIM: Write a program to perform Sentiment Analysis

Code:

```
library(tidyverse)
library(SnowballC)
library(tm)
df<-read.csv("Sentiment.csv")
summary(df)
head(df)
df<-df[c(16,6)]
head(df)
str(df)
round(prop.table(table(df$sentiment)),2)
corpus <- VCorpus(VectorSource(df$text))
as.character(corpus[[1]])
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, stemDocument)
corpus <- tm_map(corpus, stripWhitespace)
as.character(corpus[[1]])
dtm <- DocumentTermMatrix(corpus)
dtm
dim(dtm)
dtm <- removeSparseTerms(dtm, 0.999)
dim(dtm)
```

```
inspect(dtm[0:10, 1:15])
freq<- sort(colSums(as.matrix(dtm)), decreasing=TRUE)
findFreqTerms(dtm, lowfreq=60)
library(ggplot2)
wf<- data.frame(word=names(freq), freq=freq)
head(wf)
```

OUTPUT:



```
> df<-read.csv("Sentiment.csv")
> summary(df)
       id              candidate         candidate_confidence relevant_yn
 Min.   :    1   Length:13871       Min.   :0.2222       Length:13871
 1st Qu.: 3468   Class :character   1st Qu.:0.6742       Class :character
 Median : 6936   Mode  :character   Median :1.0000       Mode  :character
 Mean   : 6936                      Mean   :0.8557
 3rd Qu.:10404                      3rd Qu.:1.0000
 Max.   :13871                      Max.   :1.0000
 relevant_yn_confidence  sentiment          sentiment_confidence subject_matter
 Min.   :0.3333         Length:13871       Min.   :0.1860       Length:13871
 1st Qu.:1.0000         Class :character   1st Qu.:0.6517       Class :character
 Median :1.0000         Mode  :character   Median :0.6813       Mode  :character
 Mean   :0.9273                            Mean   :0.7569
 3rd Qu.:1.0000                            3rd Qu.:1.0000
 Max.   :1.0000                            Max.   :1.0000
 subject_matter_confidence candidate_gold         name           relevant_yn_gold
 Min.   :0.2222            Length:13871       Length:13871       Length:13871
 1st Qu.:0.6413            Class :character   Class :character   Class :character
 Median :1.0000            Mode  :character   Mode  :character   Mode  :character
 Mean   :0.7828
 3rd Qu.:1.0000
 Max.   :1.0000
 retweet_count    sentiment_gold     subject_matter_gold     text
 Min.   :   0.0   Length:13871       Length:13871        Length:13871
 1st Qu.:   0.0   Class :character   Class :character    Class :character
 Median :   2.0   Mode  :character   Mode  :character    Mode  :character
 Mean   :  45.8
 3rd Qu.:  44.0
 Max.   :4965.0
 tweet_coord        tweet_created         tweet_id        tweet_location
 Length:13871       Length:13871       Min.   :6.295e+17   Length:13871
 Class :character   Class :character   1st Qu.:6.295e+17   Class :character
 Mode  :character   Mode  :character   Median :6.297e+17   Mode  :character
                                       Mean   :6.296e+17
                                       3rd Qu.:6.297e+17
                                       Max.   :6.297e+17
 user_timezone
 Length:13871
 Class :character
 Mode  :character


> head(df)
  id             candidate candidate_confidence relevant_yn relevant_yn_confidence
1  1 No candidate mentioned               1.0000         yes                     1
2  2          Scott Walker               1.0000         yes                     1
3  3 No candidate mentioned               1.0000         yes                     1
4  4 No candidate mentioned               1.0000         yes                     1
5  5           Donald Trump              1.0000         yes                     1
6  6               Ted Cruz              0.6332         yes                     1
  sentiment sentiment_confidence    subject_matter subject_matter_confidence
1   Neutral              0.6578 None of the above                    1.0000
2  Positive              0.6333 None of the above                    1.0000
3   Neutral              0.6629 None of the above                    0.6629
4  Positive              1.0000 None of the above                    0.7039
5  Positive              0.7045 None of the above                    1.0000
6  Positive              0.6332 None of the above                    1.0000
  candidate_gold           name relevant_yn_gold retweet_count sentiment_gold
1                    I_Am_Kenzi                              5
2                 PeacefulQuest                             26
3                 PussssyCroook                             27
4               MattFromTexas31                            138
5                    sharonDay$                            156
6                  DRJohnson11                             228
  subject_matter_gold
1
2
3
4
5
```

```
 text
1                                    RT @NancyLeeGrahn: How did everyone feel about the Climate Change question la
ebate
2 RT @ScottWalker: Didn't catch the full #GOPdebate last night. Here are some of Scott's best lines in 90 seconds. 
fFâ€¦
3                                                          RT @TJMShow: No mention of Tamir Rice and the #GOPDebat
 Wow.
4 RT @RobGeorge: That Carly Fiorina is trending -- hours after HER debate -- above any of the men in just-completed
n â€¦
5 RT @DanScavino: #GOPDebate w/ @realDonaldTrump delivered the highest ratings in the history of presidential debat
coâ€¦
6        RT @GregAbbott_TX: @TedCruz: "On my first day I will rescind every illegal executive action taken by Barac
xNews
  tweet_coord             tweet_created     tweet_id tweet_location
1             2015-08-07 09:54:46 -0700 6.296972e+17
2             2015-08-07 09:54:46 -0700 6.296972e+17
3             2015-08-07 09:54:46 -0700 6.296972e+17
4             2015-08-07 09:54:45 -0700 6.296972e+17          Texas
5             2015-08-07 09:54:45 -0700 6.296972e+17
6             2015-08-07 09:54:44 -0700 6.296972e+17
               user_timezone
1                      Quito
2
3
4 Central Time (US & Canada)
5                    Arizona
6 Central Time (US & Canada)
> df<-df[c(16,6)]
> head(df)

 text
1                                    RT @NancyLeeGrahn: How did everyone feel about the Climate Change question la
ebate
2 RT @ScottWalker: Didn't catch the full #GOPdebate last night. Here are some of Scott's best lines in 90 seconds. 
fFâ€¦
3                                                          RT @TJMShow: No mention of Tamir Rice and the #GOPDebat
 Wow.
4 RT @RobGeorge: That Carly Fiorina is trending -- hours after HER debate -- above any of the men in just-completed
n â€¦
5 RT @DanScavino: #GOPDebate w/ @realDonaldTrump delivered the highest ratings in the history of presidential debat
coâ€¦
6        RT @GregAbbott_TX: @TedCruz: "On my first day I will rescind every illegal executive action taken by Barac
xNews
  sentiment
1   Neutral
2  Positive
3   Neutral
4  Positive
5  Positive
6  Positive
> str(df)
'data.frame':   13871 obs. of  2 variables:
 $ text     : chr  "RT @NancyLeeGrahn: How did everyone feel about the Climate Change question last night? Exactly.
alker: Didn't catch the full #GOPdebate last night. Here are some of Scott's best lines in 90 seconds"| __truncated
ion of Tamir Rice and the #GOPDebate was held in Cleveland? Wow." "RT @RobGeorge: That Carly Fiorina is trending --
 -- above any of the men in just-complet"| __truncated__ ...
 $ sentiment: chr  "Neutral" "Positive" "Neutral" "Positive" ...
> round(prop.table(table(df$sentiment)),2)

Negative  Neutral Positive
    0.61     0.23     0.16
> corpus <- VCorpus(VectorSource(df$text))
> as.character(corpus[[1]])
[1] "RT @NancyLeeGrahn: How did everyone feel about the Climate Change question last night? Exactly. #GOPDebate"
> corpus <- tm_map(corpus, content_transformer(tolower))
> corpus <- tm_map(corpus, removeNumbers)
> corpus <- tm_map(corpus, removePunctuation)
> corpus <- tm_map(corpus, removeWords, stopwords("english"))
> corpus <- tm_map(corpus, stemDocument)
> corpus <- tm_map(corpus, stripWhitespace)
> as.character(corpus[[1]])
```

```
> as.character(corpus[[1]])
[1] "rt nancyleegrahn everyon feel climat chang question last night exact gopdeb"
> dtm <- DocumentTermMatrix(corpus)
> dtm
<<DocumentTermMatrix (documents: 13871, terms: 16269)>>
Non-/sparse entries: 141145/225526154
Sparsity           : 100%
Maximal term length: 156
Weighting          : term frequency (tf)
> dim(dtm)
[1] 13871 16269
> dtm <- removeSparseTerms(dtm, 0.999)
> dim(dtm)
[1] 13871  1218
> inspect(dtm[0:10, 1:15])
<<DocumentTermMatrix (documents: 10, terms: 15)>>
Non-/sparse entries: 1/149
Sparsity           : 99%
Maximal term length: 7
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs â€¦ â€" â€œ aâ€¦ abc abl abort absolut abt accept
  1    0   0   0    0   0   0     0       0   0      0
  10   0   0   0    0   0   0     0       0   0      0
  2    0   0   0    0   0   0     0       0   0      0
  3    0   0   0    0   0   0     0       0   0      0
  4    1   0   0    0   0   0     0       0   0      0
  5    0   0   0    0   0   0     0       0   0      0
  6    0   0   0    0   0   0     0       0   0      0
  7    0   0   0    0   0   0     0       0   0      0
  8    0   0   0    0   0   0     0       0   0      0
  9    0   0   0    0   0   0     0       0   0      0
> freq<- sort(colSums(as.matrix(dtm)), decreasing=TRUE)
> findFreqTerms(dtm, lowfreq=60)
  [1] "â€¦"              "abort"           "act"             "actual"
  [5] "admit"           "adult"           "agre"            "alway"
  [9] "america"         "american"        "amp"             "anoth"
 [13] "answer"          "anyon"           "ask"             "attack"
 [17] "audienc"         "away"            "back"            "bad"
 [21] "balanc"          "band"            "believ"          "ben"
 [25] "berni"           "best"            "better"          "bettyfckinwhit"
 [29] "big"             "biggest"         "black"           "blacklivesmatt"
 [33] "bodi"            "boy"             "break"           "bretbaier"
 [37] "bring"           "bush"            "call"            "cam"
 [41] "came"            "campaign"        "can"             "candid"
 [45] "cant"            "car"             "care"            "carlyfiorina"
 [49] "carson"          "chang"           "check"           "cherri"
 [53] "chris"           "chrischristi"    "christi"         "christian"
 [57] "clear"           "climat"          "clinton"         "close"
 [61] "cnn"             "come"            "comment"         "conduct"
 [65] "conserv"         "control"         "correct"         "countri"
 [69] "cruz"            "cut"             "day"             "dear"
 [73] "debat"           "democrat"        "democraticdeb"   "deserv"
 [77] "didnt"           "disappoint"      "discuss"         "doesnt"
 [81] "donald"          "donaldtrump"     "donniewahlberg"  "dont"
 [85] "doubledigit"     "drink"           "ðÿ‡°ðÿ‡¸"         "elect"
 [89] "elev"            "els"             "end"             "enjoy"
 [93] "enough"          "entertain"       "ericstonestreet" "even"
 [97] "ever"            "everi"           "everyon"         "expos"
[101] "face"            "fact"            "fail"            "fair"
[105] "favorit"         "feel"            "field"           "fight"
[109] "financ"          "fiorina"         "first"           "focus"
[113] "follow"          "forward"         "fox"             "foxdeb"
[117] "foxnew"          "frankluntz"      "friend"          "frontrunn"
[121] "fuck"            "fun"             "futur"           "gâ€¦"
[125] "game"            "gay"             "get"             "give"
[129] "given"           "god"             "goldietaylor"    "good"
[133] "gop"             "gopdâ€¦"          "gopdeb"          "gopdebatâ€¦"
[137] "gopdebateâ€¦"      "got"             "govchristi"      "govern"
[141] "govmikehuckabe"  "great"           "gun"             "guy"
[145] "hair"            "hand"            "happen"          "hard"
```

```
[149] "nate"            "nead"            "near"            "nell"
[153] "help"            "hes"             "hey"             "hillari"
[157] "hillaryclinton"  "hold"            "hope"            "httpâ€¦"
[161] "httptâ€¦"         "huckabe"         "hug"             "ignor"
[165] "ill"             "illeg"           "immigr"          "import"
[169] "influenc"        "interest"        "iran"            "isnt"
[173] "issu"            "ive"             "jamiaw"          "jeb"
[177] "jebbush"         "job"             "john"            "johnkasich"
[181] "just"            "kasich"          "keep"            "kelli"
[185] "kill"            "kkkorgop"        "know"            "larryeld"
[189] "last"            "law"             "lead"            "leader"
[193] "learn"           "legitim"         "let"             "liber"
[197] "lie"             "life"            "like"            "line"
[201] "listen"          "littl"           "live"            "lol"
[205] "long"            "look"            "loser"           "lot"
[209] "love"            "lrihendri"       "made"            "make"
[213] "makeup"          "man"             "mani"            "marco"
[217] "marcorubio"      "mean"            "media"           "megyn"
[221] "megynkelli"      "men"             "mention"         "mike"
[225] "militari"        "misogyni"        "miss"            "moder"
[229] "moment"          "monaeltahawi"    "money"           "much"
[233] "music"           "nail"            "name"            "nation"
[237] "need"            "never"           "new"             "news"
[241] "next"            "night"           "nine"            "nomin"
[245] "noth"            "now"             "obama"           "obvious"
[249] "one"             "order"           "part"            "parti"
[253] "paul"            "peopl"           "perform"         "person"
[257] "pick"            "picturesshould"  "plan"            "play"
[261] "pleas"           "point"           "polici"          "polit"
[265] "politican"       "politician"      "poll"            "pose"
[269] "potus"           "presid"          "presidenti"      "pretti"
[273] "primari"         "problem"         "punch"           "purpos"
[277] "put"             "question"        "race"            "rais"
[281] "rand"            "randpaul"        "rate"            "read"
[285] "reagan"          "real"            "realbencarson"   "realdonaldtrump"
[289] "realli"          "record"          "remind"          "republican"
[293] "respect"         "rid"             "right"           "rubio"
[297] "run"             "rwsurfergirl"    "said"            "say"
[301] "scott"           "scottwalk"       "see"             "seem"
[305] "seen"            "serious"         "set"             "shit"
[309] "show"            "social"          "softbal"         "someon"
[313] "someth"          "sound"           "speak"           "stage"
[317] "stand"           "start"           "state"           "statement"
[321] "still"           "stop"            "stope"           "supermanhotmal"
[325] "support"         "sure"            "surpris"         "take"
[329] "talk"            "tax"             "tcot"            "ted"
[333] "tedcruz"         "tell"            "terror"          "thank"
[337] "that"            "thing"           "think"           "thought"
[341] "time"            "today"           "togeth"          "tonight"
[345] "top"             "total"           "tri"             "true"
[349] "trump"           "truth"           "tweet"           "twitter"
[353] "two"             "unit"            "use"             "via"
[357] "vote"            "voter"           "wait"            "walker"
[361] "wallac"          "want"            "war"             "wasnt"
[365] "watch"           "way"             "well"            "white"
[369] "will"            "win"             "winner"          "woman"
[373] "women"           "won"             "wonder"          "wont"
[377] "word"            "work"            "world"           "year"
[381] "yes"             "your"
> library(ggplot2)
> wf<- data.frame(word=names(freq), freq=freq)
> head(wf)
                      word  freq
gopdeb             gopdeb 14130
rwsurfergirl rwsurfergirl  1969
trump               trump  1964
ðÿ‡°ðÿ‡¸         ðÿ‡°ðÿ‡¸  1800
fox                   fox  1320
debat               debat  1252
>
```