Yelp Milestone Report Rubric

The goal of this task is to write a report that can be used as the basis for a final product or perhaps a submission to the Yelp dataset challenge. In addition, you need to create a 5-slide deck using RStudio Presenter tools to describe and promote your analysis.

Report

Write a 5-page report using R Markdown that describes your question/problem, how you used the dataset, the analysis you conducted, and the conclusions that you drew from the analysis. Your report must have the following sections clearly labelled:

Title - A brief description of what you've done

- Introduction - A description of the question/problem and the rationale for studying it
- Methods and Data - Describe how you used the data and the type of analytic methods that you used; it's okay to be a bit technical here but clarity is important
- Results - Describe what you found through your analysis of the data.
- Discussion - Explain how you interpret the results of your analysis and what the implications are for your question/problem.

The final report must be in PDF format using standard page sizes. Please be considerate with respect to using readable font sizes and page margins.

```
try.error = function(x)
{
   # create missing value
   y = NA
   # tryCatch error
   try_error = tryCatch(tolower(x), error=function(e) e)
   # if not an error
   if (!inherits(try_error, "error"))
   y = tolower(x)
   # result
   return(y)
}


suppressWarnings(suppressMessages(library(ggplot2)))
suppressWarnings(suppressMessages(library(xlsx)))
suppressWarnings(suppressMessages(library(sentiment)))
suppressWarnings(suppressMessages(library(wordcloud)))
suppressWarnings(suppressMessages(library(RODBC)))

setwd("D:/Google Drive/Coursera/capstone_yelt")
RndStr <- function(n = 1, lenght = 12)
{
```

```
randomString <- c(1:n) # initialize vector
for (i in 1:n)
{
randomString[i] <- paste(sample(c(0:9, letters, LETTERS),
lenght, replace = TRUE),
collapse = "")
}
return(randomString)
}
```

http://www.r-bloggers.com/using-sentiment-analysis-to-predict-ratings-of-popular-tv-series/
http://i2.wp.com/statofmind.files.wordpress.com/2014/05/rating_all_series.png
Kena buat sentimen score
Data dalam nie
C:\Program Files\R\R-3.2.1\library\sentiment\data\emotions.csv

```
conn <- odbcConnect(dsn = "capstone", uid = "hdfs", pwd = "")
sqlQuery(conn, "ADD JAR /CML/lib/lib/hive-serdes-1.0-SNAPSHOT.jar;")
```

```
## character(0)
```

```
sqlQuery(conn, "set mapred.job.priority='VERY_HIGH';")
```

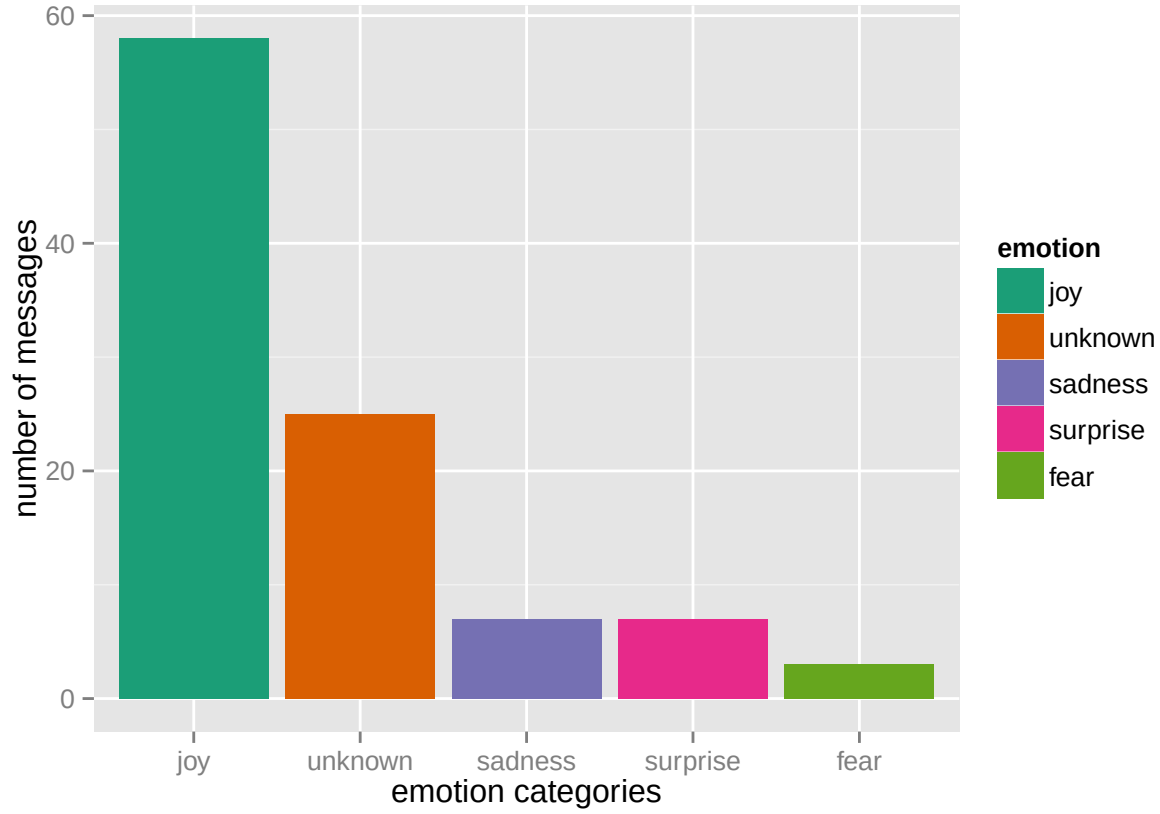
```
## character(0)
```

```
df <-
sqlQuery(
conn, "select a.user_id, a.date ,
regexp_replace(a.`text`, '\\\n|\\\r','') as review,
regexp_replace(b.`text`, '\\\n|\\\r','') as tips
from review a left join tip b on (a.business_id=b.business_id) and (a.date=b.date) and (a.user_id=b.use
where a.business_id = '4bEjOyTaDG24SY5TxsaUNQ' order by rand() limit 100")
odbcClose(conn)
head(df)
```

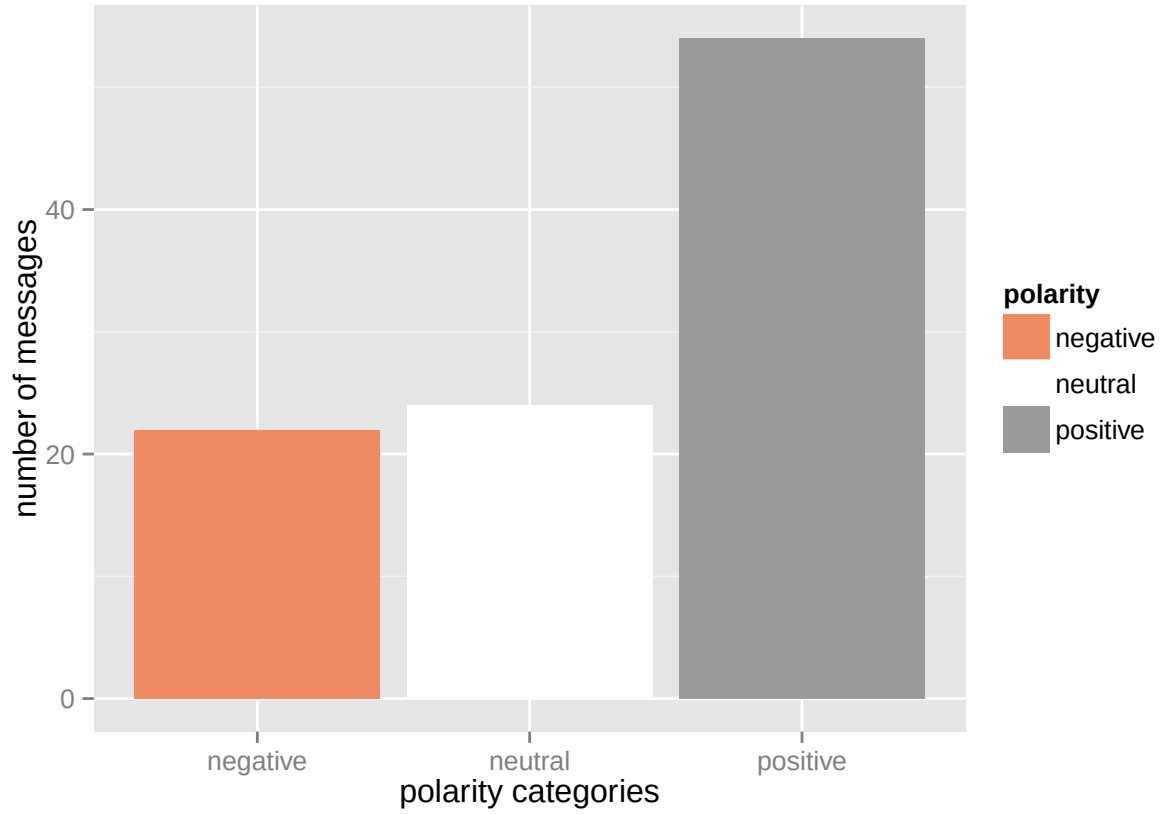
```
##                  user_id       date
## 1 Tlu_l4cgAT2jMPxUwgQVYw 2009-08-07
## 2 p6e2gwLPsFfdCebYlBgUuQ 2014-12-03
## 3 Cp2TowGXOHvyZwBxOQeN4Q 2014-06-13
## 4 97V3OvN3R4LaT91vV4dVig 2014-12-08
## 5 To2pZoDTTcJa1yC82cKdPw 2012-02-21
## 6 bEmmEYN6gq_xmuxUmCoavA 2013-11-29
##
## 1 I was really skeptical about this place.  I'm generally suspicious of chains and big restaurant co
## 2 Best place on the strip to people watch sitting on a outside table on a sunny afternoon . I had ch
## 3
## 4 Second day in LV, we started with a nice brunch at Mon Ami Gabi @ Paris Hotel. The patio was fully
## 5                                          This place is fantastic! We ate here twice while vacationi
## 6                                                                                Came here for bre
##   tips
## 1   NA
## 2   NA
```

```
## 3   NA
## 4   NA
## 5   NA
## 6   NA
```

```
#write.csv(queryResult,paste(RndStr(),".csv",sep = ""))

#df <- read.xlsx("abi_tips_review.xls",1)
df <- as.data.frame(df$review)
names(df) = "review"
df <-as.data.frame(sapply(df,gsub,pattern="[[:digit:]]",replacement=""))
df <-as.data.frame(sapply(df,gsub,pattern="[[:punct:]]",replacement=""))
df <-as.data.frame(sapply(df,gsub,pattern="@\\w+",replacement=""))
df <-as.data.frame(sapply(df,gsub,pattern="^\\s+|\\s+$",replacement=""))
df <-as.data.frame(sapply(df,gsub,pattern="[ \t]{2,}",replacement=""))

df = as.data.frame(sapply(df, try.error))
df = as.data.frame(df[!is.na(df)])
names(df) = "review"

class_emo = classify_emotion(df, algorithm="bayes", prior=1.0 )
emotion = class_emo[,7]
emotion[is.na(emotion)] = "unknown"
class_pol = classify_polarity(df, algorithm="bayes")
# get polarity best fit
polarity = class_pol[,4]

# data frame with results
sent_df = data.frame(text=df, emotion=emotion,
polarity=polarity, stringsAsFactors=FALSE)

# Sorting
sent_df = within(sent_df,
  emotion <- factor(emotion, levels=names(sort(table(emotion), decreasing=TRUE))))

# plot distribution of emotions
ggplot(sent_df, aes(x=emotion)) +
    geom_bar(aes(y=..count.., fill=emotion)) +
    scale_fill_brewer(palette="Dark2") +
    labs(x="emotion categories", y="number of messages")
```




```
ggplot(sent_df, aes(x=polarity)) +
    geom_bar(aes(y=..count.., fill=polarity)) +
    scale_fill_brewer(palette="RdGy") +
    labs(x="polarity categories", y="number of messages")
```




```
emos = levels(factor(sent_df$emotion))
nemo = length(emos)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
   tmp = df$review[emotion == emos[i]]
   emo.docs[i] = paste(tmp, collapse=" ")
}

# remove stopwords
emo.docs = removeWords(emo.docs, stopwords("english"))
# create corpus
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = emos

# comparison word cloud
suppressWarnings(suppressMessages(comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
   scale = c(3,.5), random.order = FALSE, title.size = 1.5)))
```

Slide Deck

Write 5 slides using RStudio Presenter that summarizes your report for a general audience. Explain what the problem you presented is interesting and how your analysis directly addresses that problem. The slide deck should be published as a viewable HTML presentation hosted on R Pubs or GitHub and you should be able to provide a link to the presentation so that others can view it.