DA420_Project 6_MattGraham

Steps needed to get working: 1) Uninstall all versions of R newer than 4.0.5 2) Install R v4.0.5 3) Install RTools 400 4) Install packages; Rstem takes forever. Expect a full 20-30 minutes

```
# install.packages("Rstem", repos = "http://www.omegahat.net/R", type = "source")
# install.packages("devtools")
# install.packages("wordcloud")
# install.packages("SnowballC")
# install.packages("plyr")
# install.packages("ggplot2")
# install.packages("RColorBrewer")
# install.packages("tm")
# require(devtools)
# install_url("https://cran.r-project.org/src/contrib/Archive/sentiment/sentiment_0.1.tar.g
z")
# install_url("https://cran.r-project.org/src/contrib/Archive/sentiment/sentiment_0.2.tar.g
z")
```

Import libraries

```
library(plyr)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(tm)
library(SnowballC)
library(sentiment)
```

Getting the data

```
###Get the data
data <- readLines("http://www.r-bloggers.com/wp-content/uploads/2016/01/vent.txt")
# from: http://www.wvgazettemail.com/
df <- data.frame(data)
textdata <- df$data
```

Remove non-essential characters

```
textdata = gsub("[[:punct:]]", "", textdata)
textdata = gsub("[[:digit:]]", "", textdata)
textdata = gsub("http\\w+", "", textdata)
textdata = gsub("[ \t]{2,}", "", textdata)
textdata = gsub("^\\s+|\\s+$", "", textdata)

try.error = function(x)
{
   y = NA
   try_error = tryCatch(tolower(x), error=function(e) e)
   if (!inherits(try_error, "error"))
      y = tolower(x)
   return(y)
}

textdata = sapply(textdata, try.error)
textdata = textdata[!is.na(textdata)]
names(textdata) = NULL
```

Perform sentiment analysis

```
class_emo = classify_emotion(textdata, algorithm="bayes", prior=1.0)
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom functions are
## ignored
```

```
emotion = class_emo[,7]
emotion[is.na(emotion)] = "unknown"
class_pol = classify_polarity(textdata, algorithm="bayes")
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom functions are
## ignored
```

```
polarity = class_pol[,4]

sent_df = data.frame(text=textdata, emotion=emotion,
                     polarity=polarity, stringsAsFactors=FALSE)
sent_df = within(sent_df,
                 emotion <- factor(emotion, levels=names(sort(table(emotion),
decreasing=TRUE))))
sent_df$polarity
```

```
##  [1] "positive" "negative" "negative" "negative" "neutral"  "negative"
##  [7] "negative" "positive" "positive" "negative" "positive" "positive"
## [13] "neutral"  "negative" "positive" "negative" "positive" "positive"
## [19] "positive" "positive" "positive" "positive" "neutral"  "positive"
## [25] "positive" "negative" "positive" "neutral"  "positive" "positive"
## [31] "positive" "negative" "negative" "positive" "negative" "negative"
## [37] "positive" "neutral"  "negative" "negative" "positive" "neutral"
## [43] "positive" "neutral"  "neutral"  "negative" "neutral"  "positive"
## [49] "positive" "negative" "positive" "positive" "positive" "negative"
## [55] "negative" "negative" "positive" "positive" "positive" "negative"
## [61] "positive" "positive" "negative" "negative" "positive" "positive"
## [67] "negative" "positive" "negative" "negative" "positive" "neutral"
## [73] "negative" "positive" "positive" "positive" "negative" "negative"
## [79] "negative" "neutral"  "positive" "positive" "positive"
```

```
sent_df$emotion
```

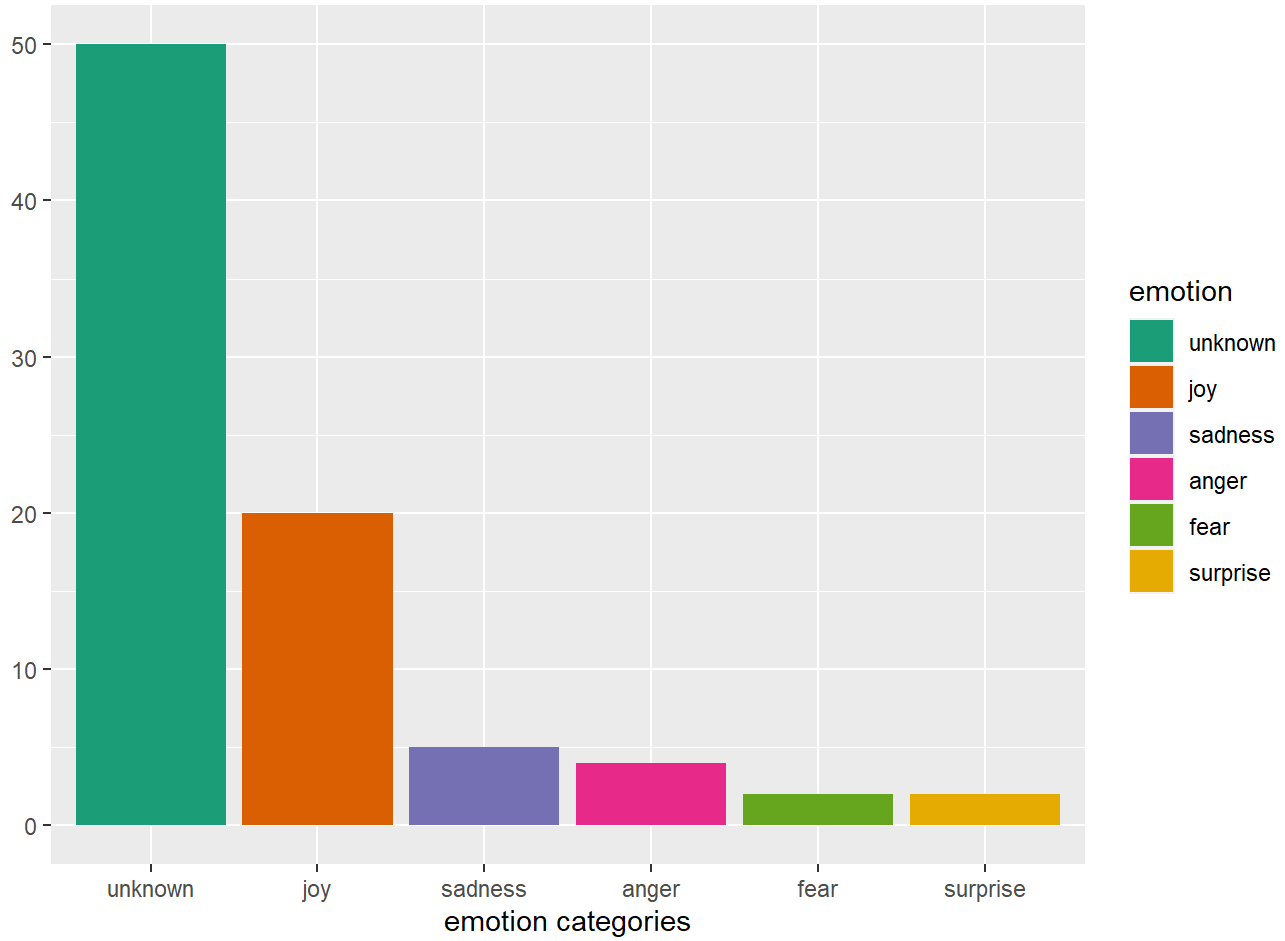
```
##  [1] unknown  unknown  anger    unknown  unknown  unknown  joy      joy
##  [9] joy      unknown  unknown  unknown  unknown  unknown  unknown  sadness
## [17] unknown  unknown  unknown  joy      unknown  unknown  unknown  unknown
## [25] unknown  sadness  surprise joy      unknown  joy      joy      sadness
## [33] unknown  unknown  sadness  joy      surprise fear     unknown  unknown
## [41] unknown  anger    unknown  unknown  unknown  anger    unknown  unknown
## [49] unknown  unknown  joy      joy      joy      anger    unknown  unknown
## [57] unknown  unknown  unknown  fear     unknown  unknown  joy      joy
## [65] unknown  unknown  joy      unknown  joy      joy      unknown  unknown
## [73] unknown  unknown  joy      unknown  unknown  unknown  sadness  joy
## [81] joy      joy      unknown
## Levels: unknown joy sadness anger fear surprise
```

Plotting our results

Emotion category plot

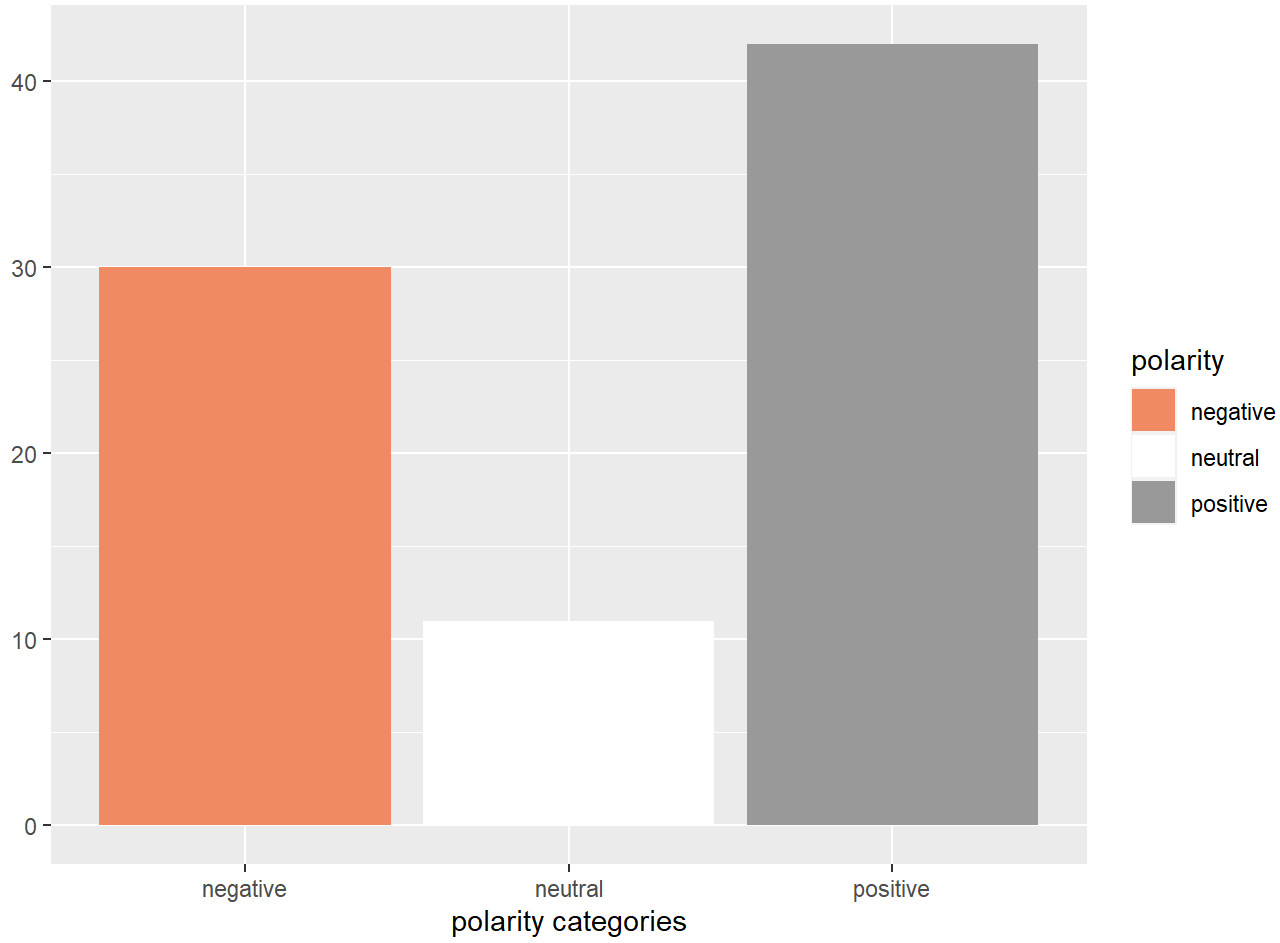
```
ggplot(sent_df, aes(x=emotion)) +
  geom_bar(aes(y=..count.., fill=emotion)) +
  scale_fill_brewer(palette="Dark2") +
  labs(x="emotion categories", y="")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```



Polarity plot

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



Word cloud

```
emos = levels(factor(sent_df$emotion))
nemo = length(emos)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
   tmp = textdata[emotion == emos[i]]
   emo.docs[i] = paste(tmp, collapse=" ")
}
emo.docs = removeWords(emo.docs, stopwords("english"))
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = emos
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
                 scale = c(3,.5), random.order = FALSE,
                 title.size = 1.5)
```

Analysis

It took a bit to get this working, but I'm so glad I got it right! I left the steps above in case other people experience this problem in the future.

Most comments are neutral, and a majority are neutral or positive. Relating this to our word cloud, we can see the proportion of positive/negative reflected with negative relating to anger and sadness. surprise and joy are likely attributed with happy, and fear/unknown being neutral.

We can see many words are filed under, "unknown". This can be something addressed and excluded in future visualizations. Or, ideally, revisited and look at things as phrases. Something like, "around" isn't a word of surprise to me. Heart under anger makes no sense either.

It would be helpful improving the bucketing algorithm. I would like to know why holiday is under surprise and not joy. Why is "aep" part of surprise? Why is it so prominent? "Don't" under joy is also a bit strange.