Project1_zz2445

Jessica Zhang

2/5/2018

```
packages.used <- c("ggplot2", "dplyr", "tibble", "tidyr",  "stringr", "tidytext", "topicmodels", "wordcl

# check packages that need to be installed.
packages.needed <- setdiff(packages.used, intersect(installed.packages()[,1], packages.used))

# install additional packages
if(length(packages.needed) > 0) {
  install.packages(packages.needed, dependencies = TRUE, repos = 'http://cran.us.r-project.org')
}

library(ggplot2)
library(dplyr)
library(tibble)
library(tidyr)
library(stringr)
library(tidytext)
library(topicmodels)
library(wordcloud)
library(ggridges)
library(lexicon)
library(tm)

source("../lib/multiplot.R")
```

Read in the data and Data Cleaning

The following code assumes that the dataset `spooky.csv` lives in a `data` folder (and that we are inside a `docs` folder).

```
spooky <- read.csv('../data/spooky.csv', as.is = TRUE)
spooky$author <- as.factor(spooky$author)
# Drop all puctuation and transform words into lower case
spooky1 <- unnest_tokens(spooky, word, text)
# Make a table with one word per row and remove `stop words`
spooky_wrd <- anti_join(spooky1, stop_words, by = "word")
```

Analysis on Authors' Uses of Discourse Markers

```
# Find discourse markers used in text
spooky_ndm <- spooky1[spooky1$word %in% discourse_markers_alemany$marker,]

# Counts total numbers of discourse markers each author used
author_dis <- count(group_by(spooky_ndm, author))
author_dis
```

```
## # A tibble: 3 x 2
## # Groups:   author [3]
##   author     n
##   <fctr> <int>
## 1    EAP 26126
## 2    HPL 20747
## 3    MWS 20995
```

```r
# Counts number of times each author used each discourse marker
author_words1 <- count(group_by(spooky_ndm, word, author))

# Counts number of times each marker was used
all_words1    <- rename(count(group_by(spooky_ndm, word)), all = n)

author_words1 <- left_join(author_words1, all_words1, by = "word")
author_words1 <- arrange(author_words1, desc(all))
author_words1 <- ungroup(head(author_words1, 81))

# Make a word cloud for Discourse Markers
words_ndm <- count(group_by(spooky_ndm, word))$word
freqs_ndm <- count(group_by(spooky_ndm, word))$n

head(sort(freqs_ndm, decreasing = TRUE))
```

```
## [1] 17956 10736  9458  6423  4347  3354
```

```r
png("../figs/Worldcloud_ndm.png")
wordcloud(words_ndm, freqs_ndm, max.words = 50, color = c("blue4", "yellow2", "grey2"))
dev.off()
```

```
## pdf
##   2
```

```r
png("../figs/ndm.png")
ggplot(author_words1) +
  geom_col(aes(reorder(word, all, FUN = min), n, fill = author)) +
  xlab(NULL) +
  coord_flip() +
  facet_wrap(~ author) +
  theme(legend.position = "none")
dev.off()
```

```
## pdf
##   2
```

Analysis on Horror Elements and Vocabulary

```r
# http://www.enchantedlearning.com/wordlist/halloween.shtml
horror <- read.csv('../data/HorrorVocab.csv', as.is = TRUE, header = FALSE)
colnames(horror) <- "vocab"

# How many horror elements or vocabulary were used in text
spooky_horror <- spooky_wrd[spooky_wrd$word %in% horror$vocab,]
png("../figs/horror.png")
ggplot(spooky_horror) +
  geom_bar(aes(author, fill = author)) +
  theme(legend.position = "none")
dev.off()
```

```
## pdf
##   2
```

```
# Wordcloud showing most frequently used horror words
words_ho <- count(group_by(spooky_horror, word))$word
freqs_ho <- count(group_by(spooky_horror, word))$n

head(sort(freqs_ho, decreasing = TRUE))
```

```
## [1] 559 380 283 240 223 203
```

```
png("../figs/Worldcloud_horror.png")
wordcloud(words_ho, freqs_ho, max.words = 50, color = c("blue4", "yellow2", "grey2"))
dev.off()
```

```
## pdf
##   2
```

Analysis with Poe's unique vocabulary list

```
# https://www.eapoe.org/papers/psblctrs/pl19741s.htm
poev <- read.csv('../data/poevocab.csv', as.is = TRUE, header = FALSE)
colnames(poev) <- "vocab"

# How many Poe's vocabulary were used in text
spooky_poe <- spooky_wrd[spooky_wrd$word %in% poev$vocab,]
png("../figs/poe.png")
ggplot(spooky_poe) +
  geom_bar(aes(author, fill = author)) +
  theme(legend.position = "none")
dev.off()
```

```
## pdf
##   2
```

```
# Wordcloud showing most frequently used Poe's words
words_poe <- count(group_by(spooky_poe, word))$word
freqs_poe <- count(group_by(spooky_poe, word))$n

head(sort(freqs_poe, decreasing = TRUE))
```

```
## [1] 23 17 13 12  9  9
```

```
png("../figs/Worldcloud_poe.png")
wordcloud(words_poe, freqs_poe, max.words = 50, color = c("blue4", "yellow2", "grey2"))
dev.off()
```

```
## pdf
##   2
```

Analysis on OED

```
# https://github.com/dwyl/english-words/blob/master/words.txt
dict <- read.csv('../data/dict.csv', as.is = TRUE, header = FALSE)
colnames(dict) <- "vocab"
dict$vocab <- tolower(dict$vocab)
spooky_oed <- spooky_wrd[!spooky_wrd$word %in% dict$vocab,]

# Wordcloud showing non-OED words
words_oed <- count(group_by(spooky_oed, word))$word
freqs_oed <- count(group_by(spooky_oed, word))$n

head(sort(freqs_oed, decreasing = TRUE))
```

```
## [1] 59 58 41 37 37 32
```

```
png("../figs/Worldcloud_oed.png")
wordcloud(words_oed, freqs_oed, max.words = 50, color = c("blue4", "yellow2", "grey2"))
dev.off()
```

```
## pdf
##   2
```

```
# Counts number of times each author used each word
author_oed <- count(group_by(spooky_oed, word, author))

# Counts number of times each word was used
all_oed  <- rename(count(group_by(spooky_oed, word)), all = n)

author_oed <- left_join(author_oed, all_oed, by = "word")
author_oed <- arrange(author_oed, desc(all))
author_oed <- ungroup(head(author_oed, 50))

png("../figs/oed.png")
ggplot(author_oed) +
  geom_col(aes(word, n, fill = author)) +
  labs(x = NULL, y = "tf-idf") +
  theme(legend.position = "none") +
  facet_wrap(~ author, ncol = 3, scales = "free") +
  coord_flip() +
  labs(y = "TF-IDF values")
dev.off()
```

```
## pdf
##   2
```

Sentiment Analysis

```
get_sentiments('nrc')
```

```
## # A tibble: 13,901 x 2
##           word sentiment
##          <chr>     <chr>
##  1     abacus     trust
##  2     abandon      fear
##  3     abandon  negative
##  4     abandon   sadness
##  5   abandoned     anger
##  6   abandoned      fear
##  7   abandoned  negative
##  8   abandoned   sadness
##  9 abandonment     anger
## 10 abandonment      fear
## # ... with 13,891 more rows
```

```r
sentiments <- inner_join(spooky_wrd, get_sentiments('nrc'), by="word")
sentiments <- rbind(sentiments[sentiments$sentiment=="negative",], sentiments[sentiments$sentiment=="fea
count(sentiments, author, sentiment)
```

```
## # A tibble: 15 x 3
##    author sentiment     n
##    <fctr>     <chr> <int>
##  1    EAP     anger  2962
##  2    EAP   disgust  2261
##  3    EAP      fear  4194
##  4    EAP  negative  7659
##  5    EAP   sadness  3938
##  6    HPL     anger  2911
##  7    HPL   disgust  2490
##  8    HPL      fear  4435
##  9    HPL  negative  7385
## 10    HPL   sadness  3571
## 11    MWS     anger  3996
## 12    MWS   disgust  2946
## 13    MWS      fear  5298
## 14    MWS  negative  8630
## 15    MWS   sadness  5165
```

```r
png("../figs/sa.png")
ggplot(count(sentiments, author, sentiment)) +
  geom_col(aes(sentiment, n, fill = sentiment)) +
  facet_wrap(~ author) +
  coord_flip() +
  theme(legend.position = "none")
dev.off()
```

```
## pdf
##   2
```

```r
nrow(sentiments[sentiments$author=="MWS",])
```

```
## [1] 26035
```

```r
nrow(sentiments[sentiments$author=="EAP",])
```

```
## [1] 21014
```

```r
nrow(sentiments[sentiments$author=="HPL",])
```

```
## [1] 20792
```