Goal

Goal of this project is to find characteristics of texts from 3 popular horror authors, identify similarities and differences in their texts in the spooky dataset. Data consists of excerpts of texts written by Edgar Allan Poe (EAP), HP Lovecraft (HPL), and Mary Wollstonecraft Shelley (MWS).

Load packages and read the data

Setup the libraries if not already installed

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

# 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(plotrix)
library(waffle)
```

Read in the data

`spooky.csv` in `data` folder, and this Rmd inside `doc` folder.

```
spooky <- read.csv('../data/spooky.csv', as.is = TRUE)
```

Overview of the dataset

Take a look of first few rows and dimension of the dataset

```
head(spooky, 3)
```

```
##        id
## 1 id26305
## 2 id17569
## 3 id11008
##
## 1 This process, however, afforded me no means of ascertaining the dimensions of my dungeon; as I mig
## 2
## 3                                  In his left hand was a gold snuff box, from which, as he capered dow
##   author
## 1    EAP
## 2    HPL
## 3    EAP
```

```
dim(spooky)
```

```
## [1] 19579     3
```

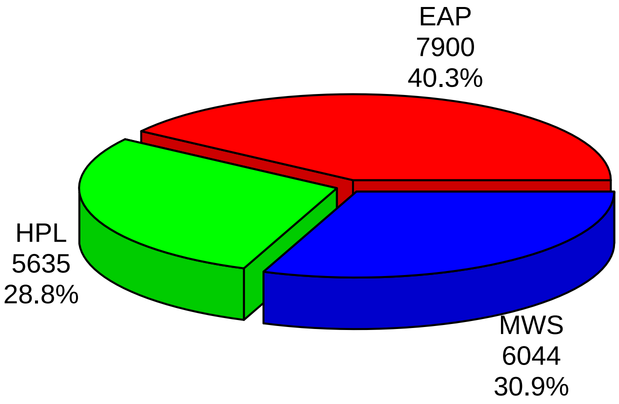
How many texts do each author have in the dataset?

```
mytable <- table(spooky$author)
mytable
```

```
##
##  EAP  HPL  MWS
## 7900 5635 6044
```

Plot composition of number of texts from 3 authors in pie chart, display counts and percentages

```
lbls <- paste(names(mytable), '\n', mytable, '\n', round(mytable/sum(mytable) * 100, 1), '%', sep = '')
pie3D(mytable, labels = lbls, explode = 0.05, labelcex = 0.8)
```



Writing Style

Do some authors use more questions in the texts than others?

- Count number of question marks in texts for spooky
- Add a field `num_qns` for the counts
- Wrangle data to show counts for each author
- Plot a waffle chart to see comparison of use of questions in texts among 3 authors.

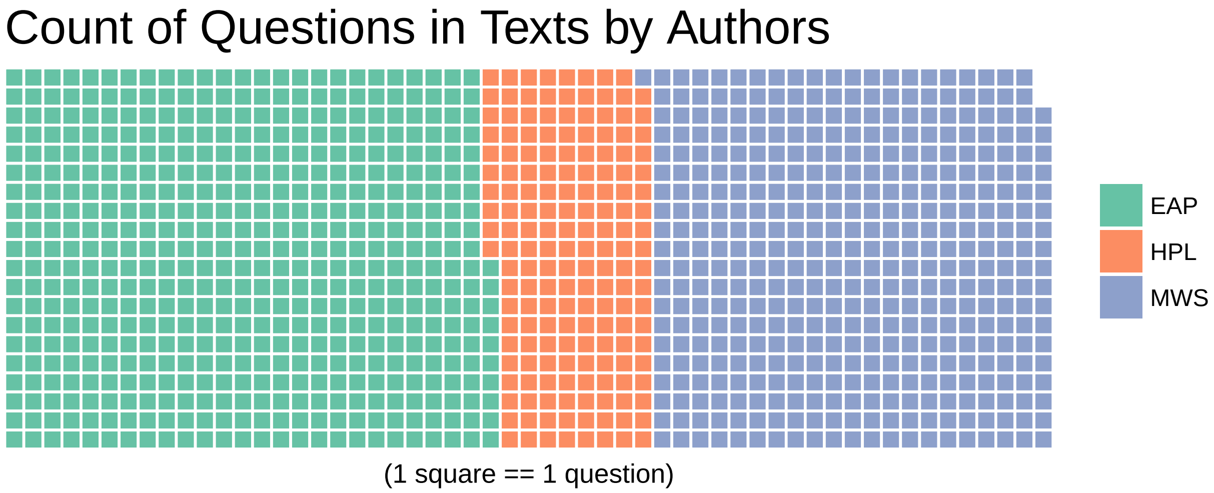
```
str_count(spooky, '\\?')
```

```
## [1]    0 1098    0
```

```
dat1 <- mutate(spooky, num_qns = str_count(spooky$text, '\\?'))
dat2 <- aggregate(dat1$num_qns, by = list(Author = dat1$author), FUN = sum)
dat2
```

```
##   Author   x
## 1    EAP 510
## 2    HPL 169
## 3    MWS 419
```

```
waffle(c('EAP' = dat2[1, 2], 'HPL' = dat2[2, 2], 'MWS' = dat2[3, 2]), rows = 20, size = 0.5, title = 'C
```



Sentiment analysis

Positive and negative emotional content comparison in authors' text

Apply sentiment analysis using bing lexicon

```
get_sentiments("bing")
```

```
## # A tibble: 6,788 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faced     negative
##  2 2-faces     negative
##  3 a+          positive
##  4 abnormal    negative
##  5 abolish     negative
##  6 abominable  negative
##  7 abominably  negative
##  8 abominate   negative
##  9 abomination negative
## 10 abort       negative
## # ... with 6,778 more rows
```

```
tidy_text <- unnest_tokens(spooky, word, text)
tidy_text_sentiment <- tidy_text %>% inner_join(get_sentiments('bing'))
```

```
## Joining, by = "word"
```

```
head(tidy_text_sentiment, 10)
```

```
##          id author        word sentiment
## 1  id26305    EAP     dungeon  negative
## 2  id26305    EAP   perfectly  positive
## 3  id17569    HPL     mistake  negative
## 4  id11008    EAP        gold  positive
## 5  id11008    EAP   fantastic  positive
## 6  id11008    EAP incessantly  negative
## 7  id11008    EAP    greatest  positive
## 8  id27763    MWS      lovely  positive
## 9  id27763    MWS     fertile  positive
## 10 id27763    MWS       happy  positive
```

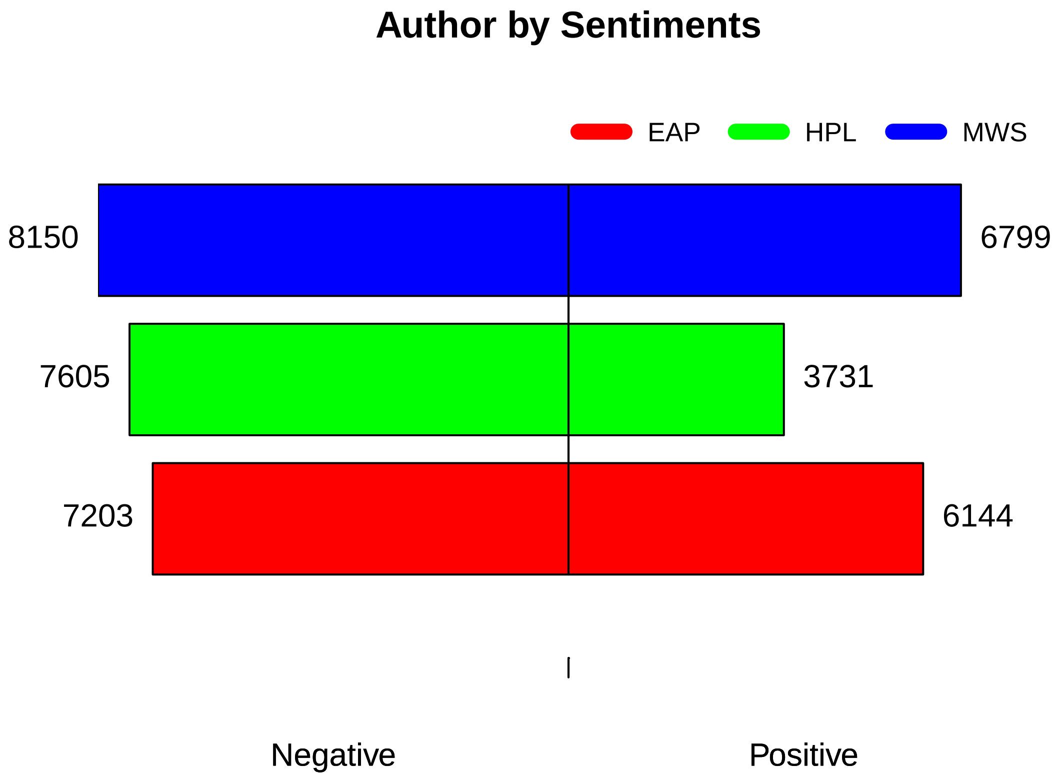
```
dat3 <- table(tidy_text_sentiment$sentiment, tidy_text_sentiment$author)
dat3
```

```
##
##             EAP  HPL  MWS
##   negative 7203 7605 8150
##   positive 6144 3731 6799
```

```
pyramid.plot(dat3[1,c(1:3)], dat3[2,c(1:3)], top.labels = NULL, show.values = TRUE, ndig = 0, main = 'A
```

```
## [1] 5.1 4.1 4.1 2.1
```

```
legend('topright', legend = c("EAP", "HPL", "MWS"), col = c("red", "green", "blue"), lty = 1, bty = 'n'
```


Author by Sentiments
EAP
HPL
MWS
Negative
Positive