Some Simple SPOOKY Data Analysis

Pak Kin (Chris) Lai

January 22, 2018

1 Introduction

This files contains some simple analysis of the SPOOKY data. The goal is to remind ourselves of some of our basic tools for working with text data in `R` and also to practice reproducibility. You should be able to put this file in the `doc` folder of your `Project 1` repository and it should just run (provided you have `multiplot.R` in the `libs` folder and `spooky.csv` in the `data` folder). If you open to file from a forked `Week1-GitHub` repo, you should have no trouble running the code directly.

2 Preparation

2.1 Library

First we want to install and load libraries we need along the way.

```
packages.used <- c("ggplot2", "dplyr", "tidytext", "wordcloud", "stringr", "ggridges", "SnowballC")

# 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, repos = 'http://cran.us.r-project.org')
}

library(ggplot2)
library(dplyr)
library(tidytext)
library(wordcloud)
library(stringr)
library(ggridges)
library(SnowballC)
```

2.2 Helper Function

The only helper function that we use is the multiplot function.

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

2.3 Data

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

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

3 Overview

3.1 Structure

The structure of the data is as follows:

```
head(spooky)
```

```
##        id
## 1 id26305
## 2 id17569
## 3 id11008
## 4 id27763
## 5 id12958
## 6 id22965
##
## 1
## 2
## 3
## 4
## 5
## 6 A youth passed in solitude, my best years spent under your gentle and feminine fosterage, has so r
##   author
## 1    EAP
## 2    HPL
## 3    EAP
## 4    MWS
## 5    HPL
## 6    MWS
```

```
summary(spooky)
```

```
##       id                text              author
##  Length:19579       Length:19579       Length:19579
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

Each row of our data contains a unique ID, a single sentence text excerpt, and an abbreviated author name. `HPL` is Lovecraft, `MWS` is Shelly, and `EAP` is Poe.

3.2 Missing Values

There are no missing values.

```
sum(is.na(spooky))
```

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

3.3 Reformatting Author

Changing the author name to be a factor variable will help us later on.

```
spooky$author <- as.factor(spooky$author)
```

4 Data Manipulation

4.1 Data Cleaning

The `unnest_tokens()` function drops all punctuation and transforms all words into lower case.

```
spooky_wrd <- unnest_tokens(spooky, word, text)
head(spooky_wrd)
```

```
##          id author     word
## 1   id26305    EAP     this
## 1.1 id26305    EAP  process
## 1.2 id26305    EAP  however
## 1.3 id26305    EAP afforded
## 1.4 id26305    EAP       me
## 1.5 id26305    EAP       no
```

4.2 Stop Words

Before filtering out the stop words, it may be useful to observe the most frequently used stop words for each author. We replicate a dataset so we can continue our analysis later. Then we filter out the stop words using tidytext's dictionary of stop words.

```
spooky_wrd_stop <- spooky_wrd
spooky_wrd <- anti_join(spooky_wrd, stop_words, by = "word")
head(spooky_wrd)
```

```
##        id author         word
## 1 id26305    EAP      process
## 2 id26305    EAP     afforded
## 3 id26305    EAP        means
## 4 id26305    EAP ascertaining
## 5 id26305    EAP   dimensions
## 6 id26305    EAP      dungeon
```

4.3 Frequency of Author

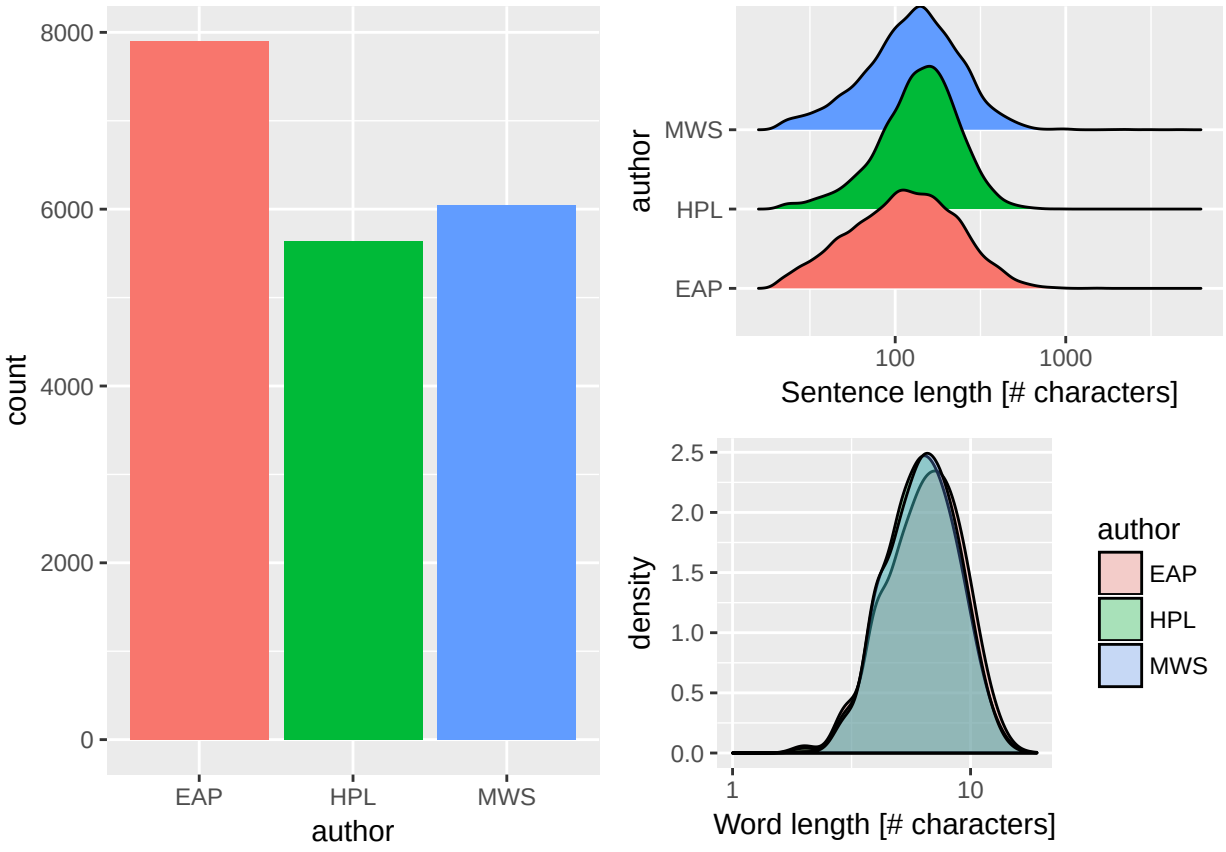
I created a dataset of the frequency of the words by each author for future analyses.

```
EAP <- sum(spooky_wrd$author == "EAP")
HPL <- sum(spooky_wrd$author == "HPL")
MWS <- sum(spooky_wrd$author == "MWS")

frequency <- data.frame(EAP, HPL, MWS, row.names = "Frequency of words")
```

5 Sentence and Structure Analyses

5.1 Total Featured, Sentence Length, Word Length

```
p1 <- ggplot(spooky) +
      geom_bar(aes(author, fill = author)) +
      theme(legend.position = "none")


spooky$sen_length <- str_length(spooky$text)

p2 <- ggplot(spooky) +
      geom_density_ridges(aes(sen_length, author, fill = author)) +
      scale_x_log10() +
      theme(legend.position = "none") +
      labs(x = "Sentence length [# characters]")


spooky_wrd$word_length <- str_length(spooky_wrd$word)

p3 <- ggplot(spooky_wrd) +
      geom_density(aes(word_length, fill = author), bw = 0.05, alpha = 0.3) +
      scale_x_log10() +
      labs(x = "Word length [# characters]")

layout <- matrix(c(1, 2, 1, 3), 2, 2, byrow = TRUE)
multiplot(p1, p2, p3, layout = layout)
```



Observations:

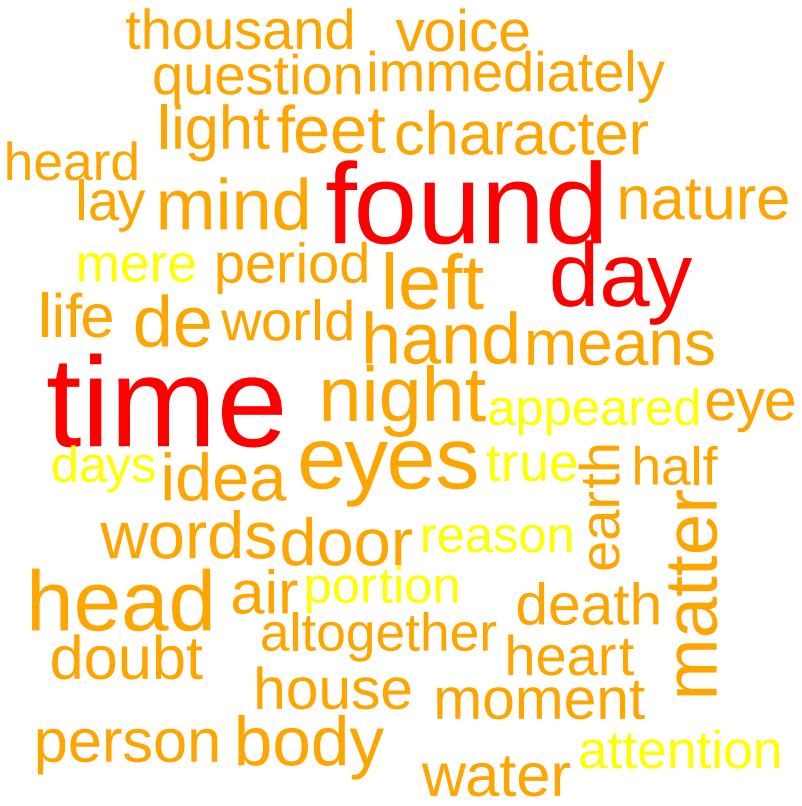
1. EAP is featured most frequently.
2. Sentence length varies more for EAP.
3. EAP has slightly longer words than the others, whereas MWS has slightly shorter words than the others.

5.2 Wordclouds

```
words_EAP <- count(group_by(spooky_wrd[spooky_wrd$author == "EAP",], word))$word
freqs_EAP <- count(group_by(spooky_wrd[spooky_wrd$author == "EAP",], word))$n

words_HPL <- count(group_by(spooky_wrd[spooky_wrd$author == "HPL",], word))$word
freqs_HPL <- count(group_by(spooky_wrd[spooky_wrd$author == "HPL",], word))$n

words_MWS <- count(group_by(spooky_wrd[spooky_wrd$author == "MWS",], word))$word
freqs_MWS <- count(group_by(spooky_wrd[spooky_wrd$author == "MWS",], word))$n


wordcloud(words_EAP, freqs_EAP, max.words = 50, color = c("yellow", "orange", "red"))
```

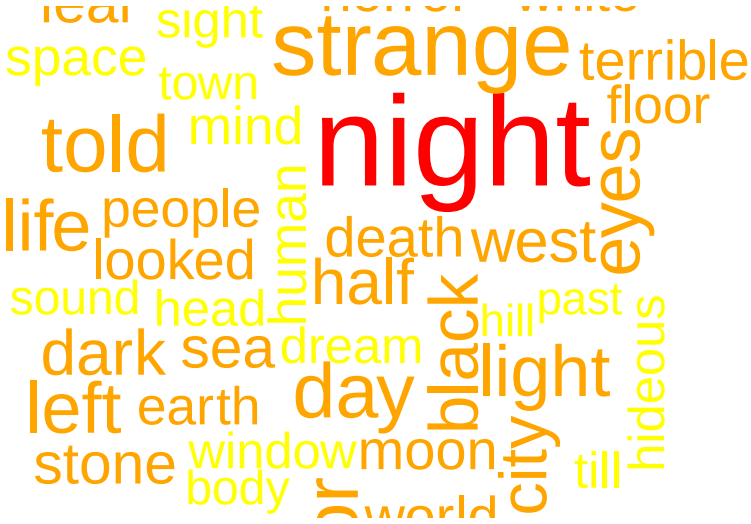
```
## Warning in wordcloud(words_EAP, freqs_EAP, max.words = 50, color =
## c("yellow", : manner could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words_EAP, freqs_EAP, max.words = 50, color =
## c("yellow", : length could not be fit on page. It will not be plotted.
```




```
wordcloud(words_HPL, freqs_HPL, max.words = 50, color = c("yellow", "orange", "red"))
```

```
## Warning in wordcloud(words_HPL, freqs_HPL, max.words = 50, color =
## c("yellow", : ancient could not be fit on page. It will not be plotted.
```




```
wordcloud(words_MWS, freqs_MWS, max.words = 50, color = c("yellow", "orange", "red"))
```

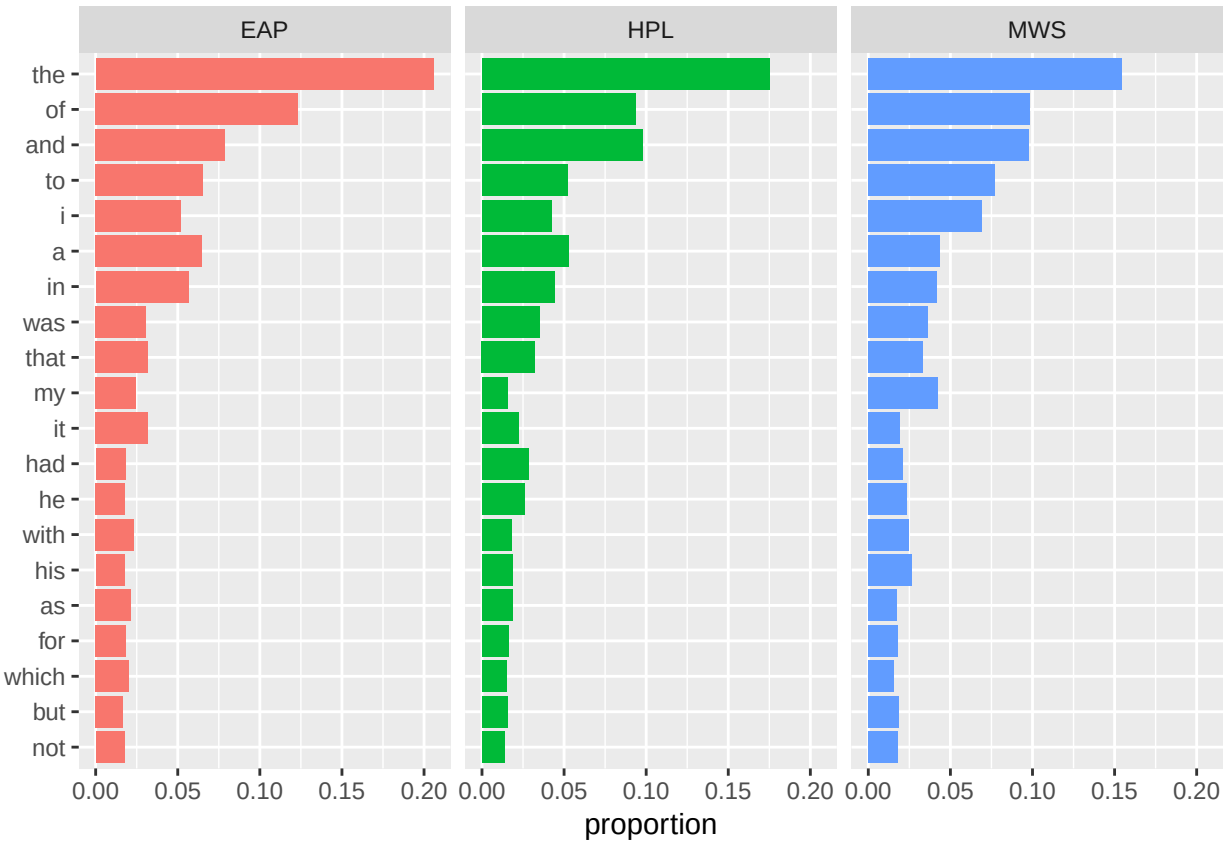
```
## Warning in wordcloud(words_MWS, freqs_MWS, max.words = 50, color =
## c("yellow", : love could not be fit on page. It will not be plotted.
```

6 Words Analyses

Stop Words Analysis

Using the dataset that includes the stop words, we can simply look at the most commonly used word grouped by each author since stop words are typically the most commonly used words. However, we look at the proportion instead of the total number of appearances since our earlier observation shows that authors are featured differently in this data.

```
# Counts number of times each author used each word.
author_words <- count(group_by(spooky_wrd_stop, word, author))

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

author_words <- left_join(author_words, all_words, by = "word")
author_words <- arrange(author_words, desc(all))
author_words <- ungroup(head(author_words, 60))

# Counts the proportion of the word in the total number of words featured by each author
author_words$proportion <- ifelse(author_words$author == "EAP", author_words$n/frequency$EAP, ifelse(au

ggplot(author_words) +
  geom_col(aes(reorder(word, all, FUN = min), proportion, fill = author)) +
  xlab(NULL) +
  coord_flip() +
  facet_wrap(~ author) +
  theme(legend.position = "none")
```



Observations:

1. EAP uses the words 'the', 'of', 'a', 'in' much more frequently than the other two authors, and "and' much less frequently than the others.
2. EAP uses the word 'of' much more frequently than he uses 'and', whereas for the other two authors, the frequency of these two words are around the same.
3. HPL uses 'had' more frequently than the other two authors, but only by a small margin.
4. MWP uses the words "I", 'my', 'his' more frequently than the other two authors, but only by a small margin.
5. HPL uses 'my' less frequently than the other two authors, but only by a small margin.

Accented Words

4.3 Stem Words

When we analyze the words later, it will be helpful to group them by their stem instead. We replicate a dataset so we can continue our analysis later.

```
spooky_wrd_stem <- spooky_wrd
spooky_wrd_stem$word <- wordStem(spooky_wrd_stem$word)
```