The Similarities and Differences among Spooky Authors

Pak Kin (Chris) Lai

February 4, 2018

1 Introduction

This project's purpose is to use various techniques to study the similarities and the differences in writing among three horror story authors, Edgar Allan Poe, HP Lovecraft, and Mary Shelley. This project's analysis is split into 3 major sections: Sentence and Structure Analyses, Characteristic Word Analyses, Pair Analyses, and Sentiment Analyses.

Sections 2-4 exist to setup the data used for the analyses. Section 5 onwards contain the actual analyses, so feel free to skip to that section if you prefer.

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", "ti

# 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(plyr)
library(dplyr)
library(tidytext)
library(wordcloud)
library(stringr)
library(ggridges)
library(SnowballC)
library(tidyr)
library(widyr)
```

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

We also see that 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

Now we want to manipulate the data so our analysis could be done easier.

4.1 General Dataset

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 for the original dataset 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 realized that authors may be featured at different amounts for our data. To compensate for this, I created a dataframe of the frequency of the words by each author.

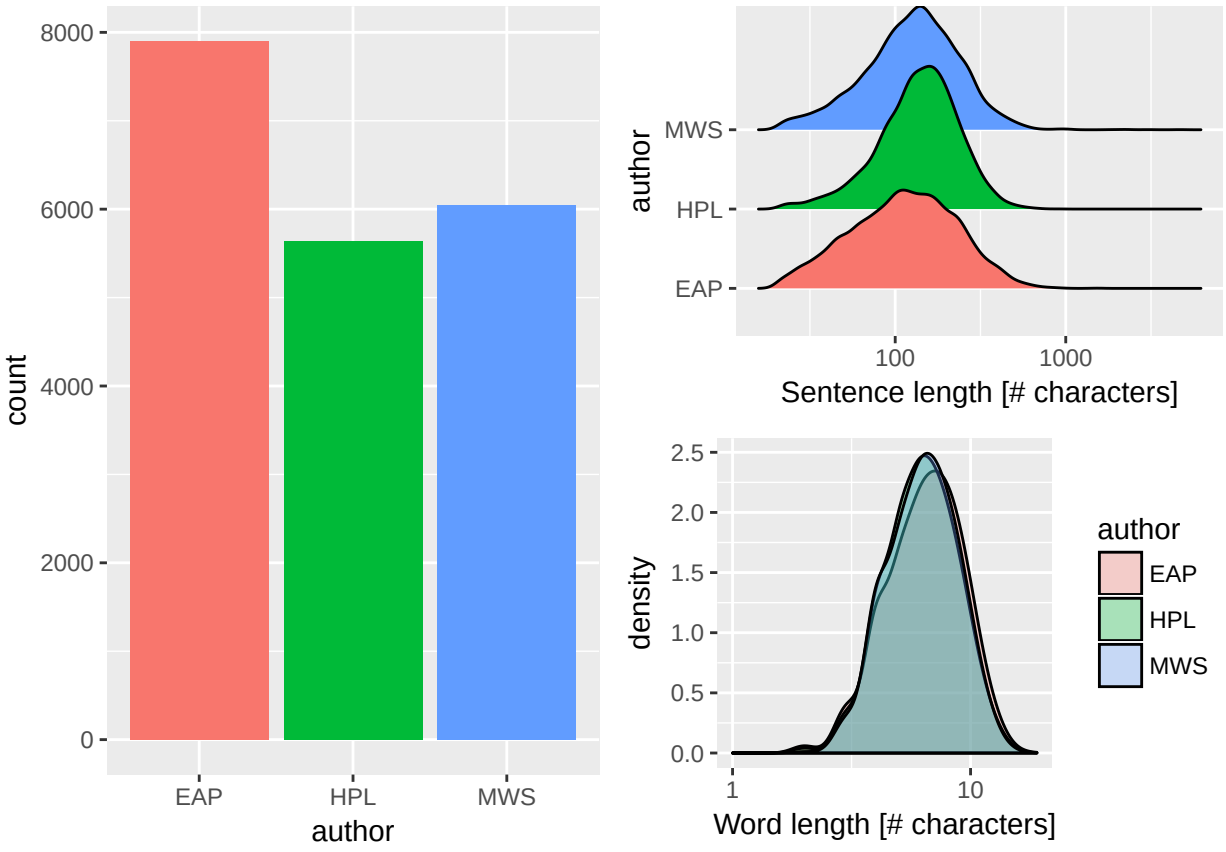
```
EAP <- sum(spooky_wrd$author == "EAP")
HPL <- sum(spooky_wrd$author == "HPL")
MWS <- sum(spooky_wrd$author == "MWS")

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

5 Sentence and Structure Analyses

This section allows us to look at the general structure of the sentences for each author.

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.

6 Characteristic Word Analyses

This section is to study the most frequently used characteristic word for each author.

6.1 Original Dataset

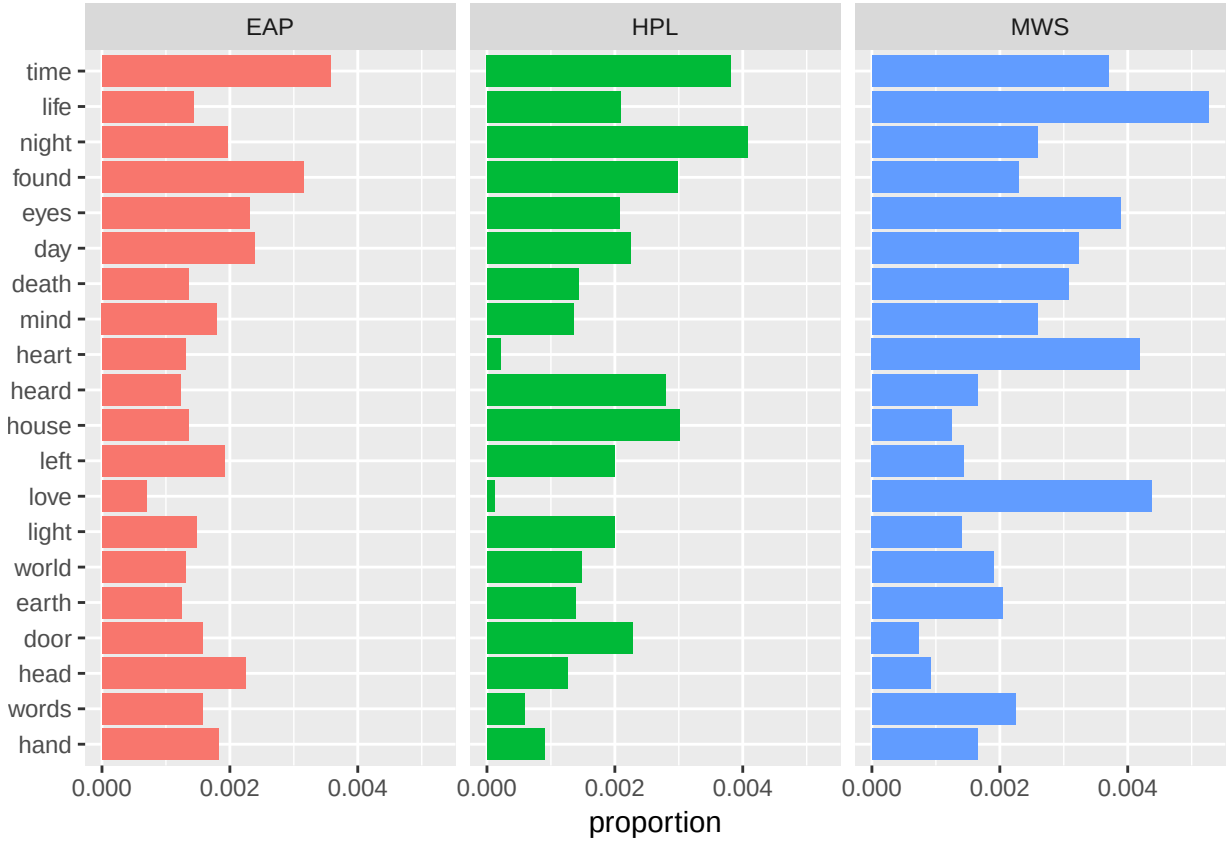
This is the analyis done on our original dataset with the lower case and without punctuation.

6.1.i Proportion of Words

We first look at the highest proportion of the characteristic words used by each author.

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

# Counts number of times each word was used.
all_words    <- rename(count(group_by(spooky_wrd, 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. 'Time' is used commonly by all three authors.
2. 'Night' is used more frequently by HPL.
3. 'Life', 'heart', 'eyes', 'day', 'death', and 'love' are used much more frequently by MWS. HPL seldom uses the word 'love'.
4. 'Heard', 'house', annd 'door' are used more frequently by HPL.
5. 'Head' is used frequently by EAP.

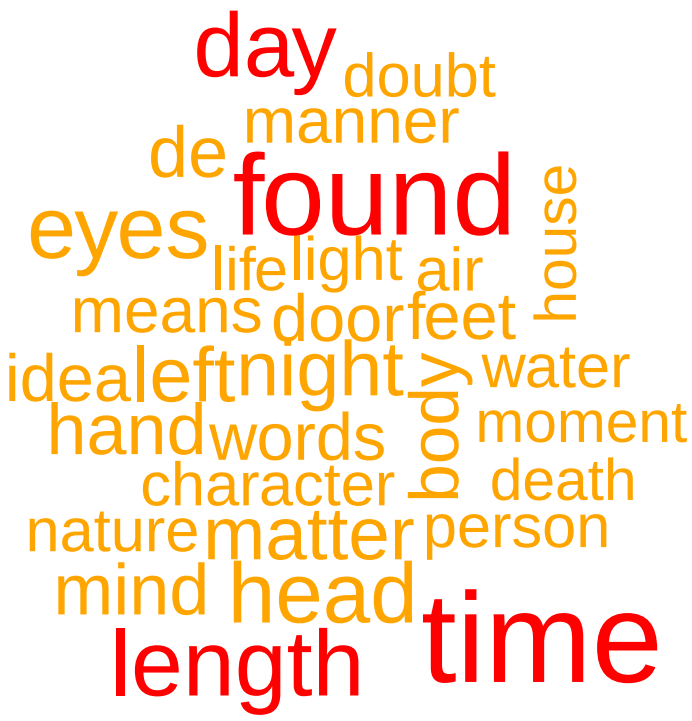
6.1.ii Word Clouds

Word Clouds may also give us an interesting illustration.

```
#Count each word grouped by author
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
```

EAP

```
wordcloud(words_EAP, freqs_EAP, max.words = 30, color = c("yellow", "orange", "red"))
```



HPL

```
wordcloud(words_HPL, freqs_HPL, max.words = 30, color = c("yellow", "orange", "red"))
```

MWS

```
wordcloud(words_MWS, freqs_MWS, max.words = 30, color = c("yellow", "orange", "red"))
```

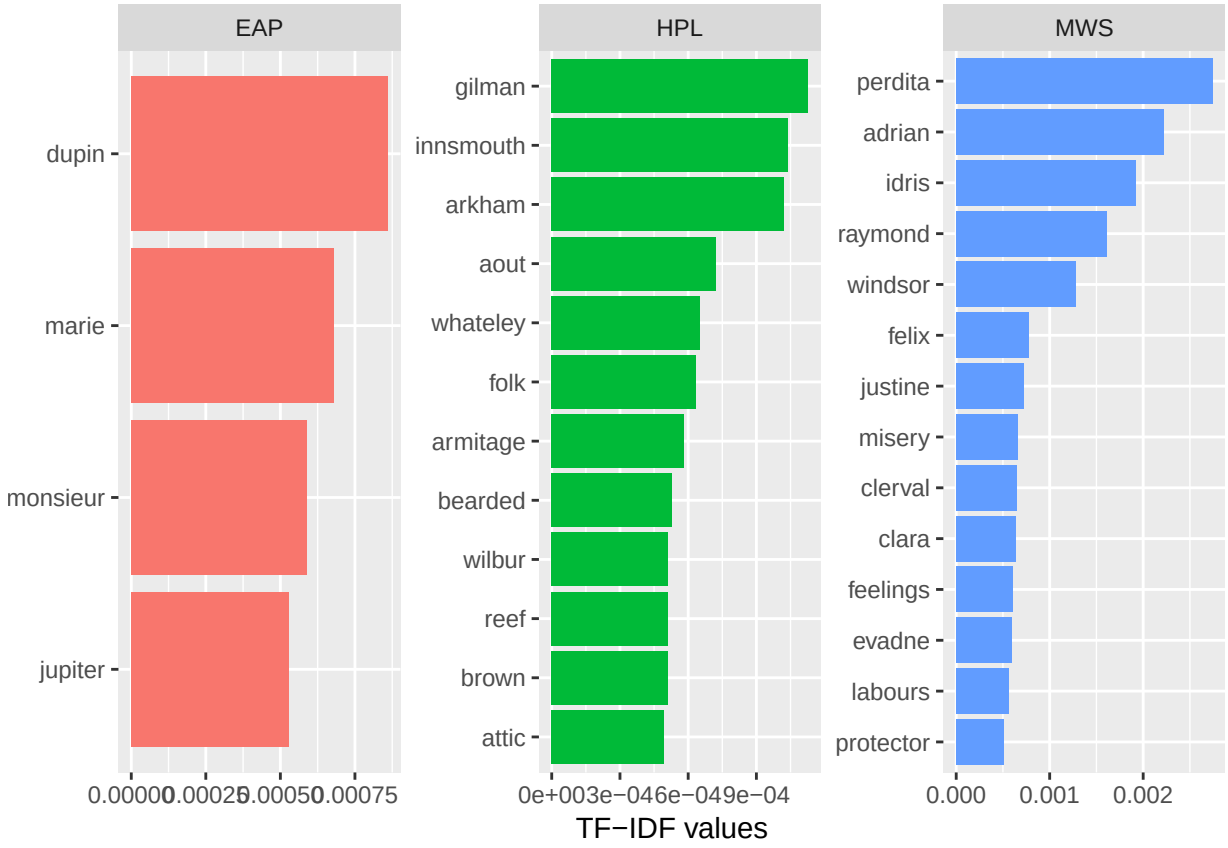
Observations:

1. 'Time' and 'Found' are very common words among the three authors.
2. EAP prefers the word 'Day', whereas HPL prefers the word 'Night'.
3. 'Love', 'Heart', 'Life' are common words for MWS.

6.1.i Tf-idf

A Tf-idf analysis will give us a measure of the frequency of a word used by one author relative to the frequency of that word used by the others. In other words, it adjusts for the rarity of the word.

```
# Calculate the tf-idf for each word
tf_idf <- spooky_wrd %>%
  count(author, word) %>%
  bind_tf_idf(word, author, n) %>%
  arrange(desc(tf_idf))

# Give the top 30 tf-idfs
tf_idf <- head(tf_idf, 30)

ggplot(tf_idf) +
  geom_col(aes(reorder(word, tf_idf), tf_idf, 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")
```



Observations:

1. As we expect, names such as 'Dupin', 'Gilman', 'Perdita' are the words with the highest tf-idf.
2. EAP's 'Monsieur' has a high tf-idf. However, she has the least amount of words in the top 30 tf-idfs shown above.
3. HPL's 'folk', 'bearded', 'reef', 'brown', and 'attic' are interesting words with high tf-idf.

4. MWS's 'misery', 'feelings', labours', and 'protector' are words with high tf-idf. They are words most associated with emotions and protections. MWS also has the most amount of words in the top 30 tf-idfs.

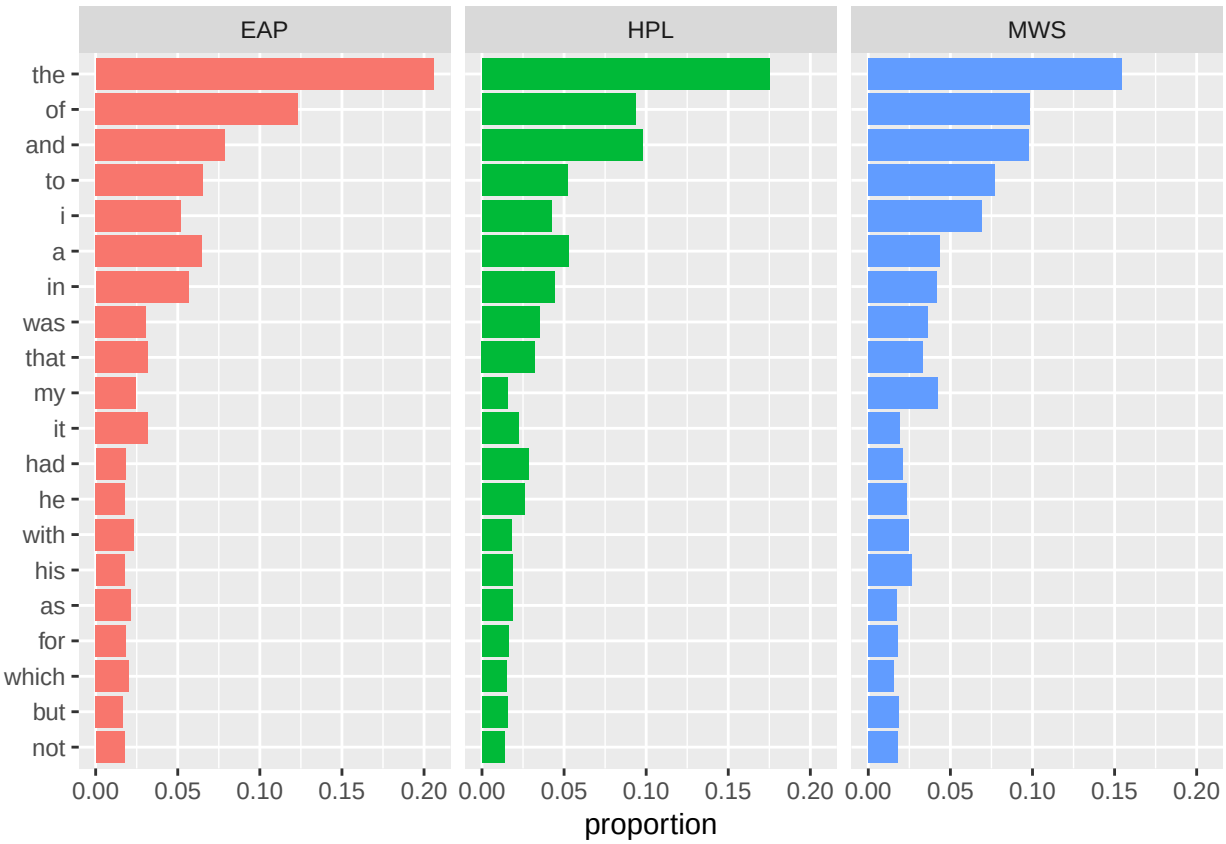
6.2 Stop Words Dataset

This is the analysis done on the stop dataset that we created earlier.

6.2.i Proportion of Words

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_stop <- count(group_by(spooky_wrd_stop, word, author))

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

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

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

ggplot(author_words_stop) +
  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.

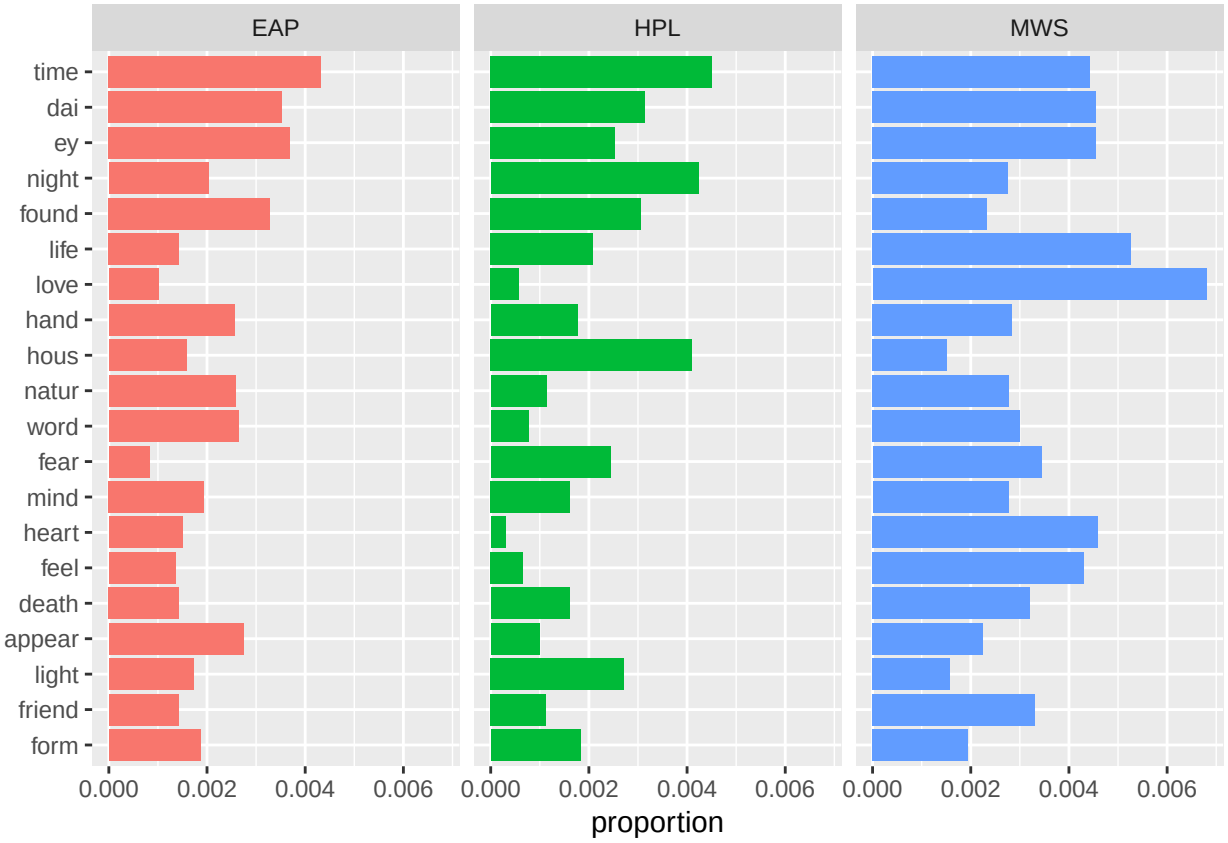
6.3 Stem Words Dataset

It might be interesting to see the words grouped by their stem instead. To do this, I duplicated the dataset and did the same analyses as above.

6.3.i Proportion of Words

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

# Counts number of times each author used each word.
author_words_stem <- count(group_by(spooky_wrd_stem, word, author))

# Counts number of times each word was used.
```

```r
all_words_stem    <- rename(count(group_by(spooky_wrd_stem, word)), all = n)

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

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

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



Observations:

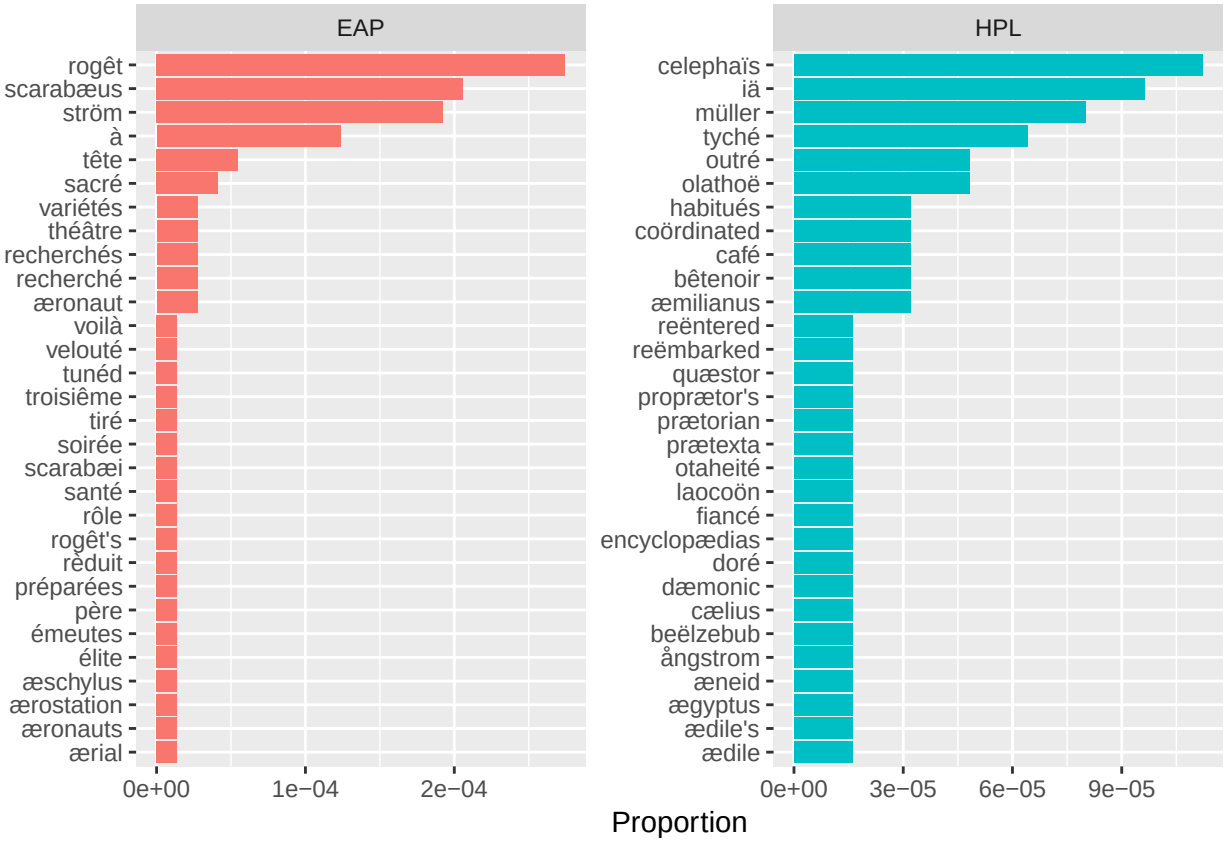
1. 'Dai', 'ey', 'heart', 'feel', and 'death' are used more frequently by MWS.
2. 'Love' is used significantly more frequently by MWS.
3. 'Natur', 'word', and 'appear' is used by EAP more frequently.
4. 'Night' and 'hous' is used much more frequently by HPL.
5. 'Fear' is used less frequently compared to the other authors.

6.4 Accented Words Dataset

6.4.i Proportion of Words

Another interesting analysis to make is to look at the frequency/proportion of accented words for each author.

```
accented <- grep('à|á|â|ä|æ|ã|å|ā|è|é|ê|ë|ē|ė|ę|î|ï|í|ī|į|ì|ô|ö|ò|ó|œ|ø|ō|õ|û|ü|ù|ú|ū', spooky_wrd$word)
spooky_wrd_accented <- spooky_wrd[accented,]

# Counts number of times each author used each word.
author_words_accented <- count(group_by(spooky_wrd_accented, word, author))
author_words_accented <- arrange(author_words_accented, desc(n))

# Gets the top 30
author_words_accented <- group_by(author_words_accented, author)
author_words_accented <- top_n(author_words_accented, 30, word)

# Counts the proportion of the word in the total number of words featured by each author
author_words_accented$proportion <- ifelse(author_words_accented$author == "EAP", author_words_accented$

ggplot(author_words_accented) +
  #geom_col(aes(reorder(word, all, FUN = min), proportion, fill = author)) +
  geom_col(aes(reorder(word, proportion), proportion, fill = author)) +
  labs(x = NULL, y = "Proportion") +
  coord_flip() +
  facet_wrap(~ author, ncol = 2, scales = "free") +
  theme(legend.position = "none")
```



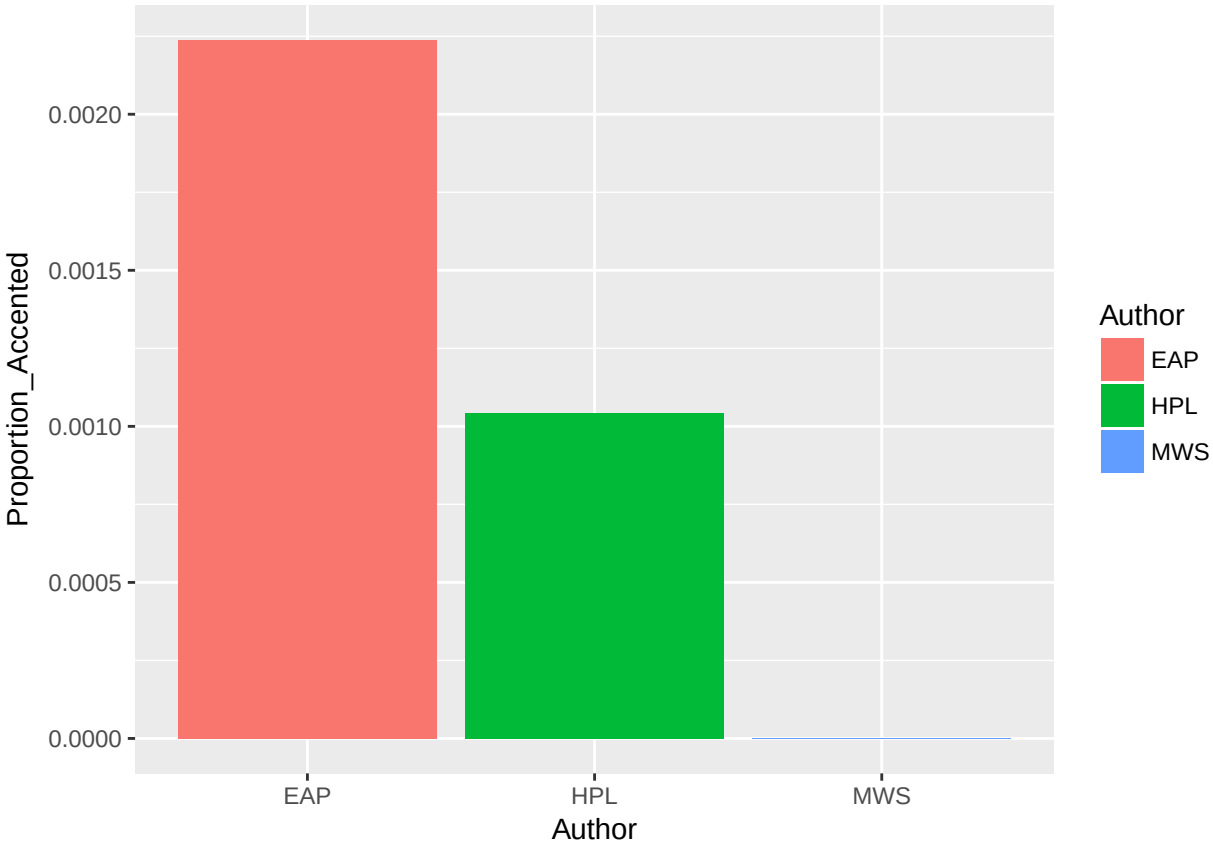
Observations:

1. We see that only EAP and HPL uses these accented words.
2. EAP uses the word 'rogeêt' a lot.
3. HPL uses the word 'celephaïs a lot.

6.4.ii Proportion of All Words

However, there isn't many overlap of the specific accented words used between the two authors. As a result, it may be more useful to just simply see the proportion of all accented words used by each author.

```
EAP_accented <- sum(spooky_wrd_accented$author == "EAP")
HPL_accented <- sum(spooky_wrd_accented$author == "HPL")
MWS_accented <- sum(spooky_wrd_accented$author == "MWS")

# Counts the proportion of the accented word in the total number of words featured by each author
accented <- data.frame(Author = c("EAP", "HPL", "MWS"), Proportion_Accented = c(EAP_accented/frequency$

ggplot(data = accented, aes(x = Author, y = Proportion_Accented, fill = Author)) +
  geom_bar(stat="identity")
```



Observations:

1. As we can see, EAP has more than double the proportion of accented words compared to that of HPL.
2. MWS doesn't use accented words at all.

7 Pair Analyses

Now we might want to move on from looking at single words to two words.

7.1 Bigram Dataset

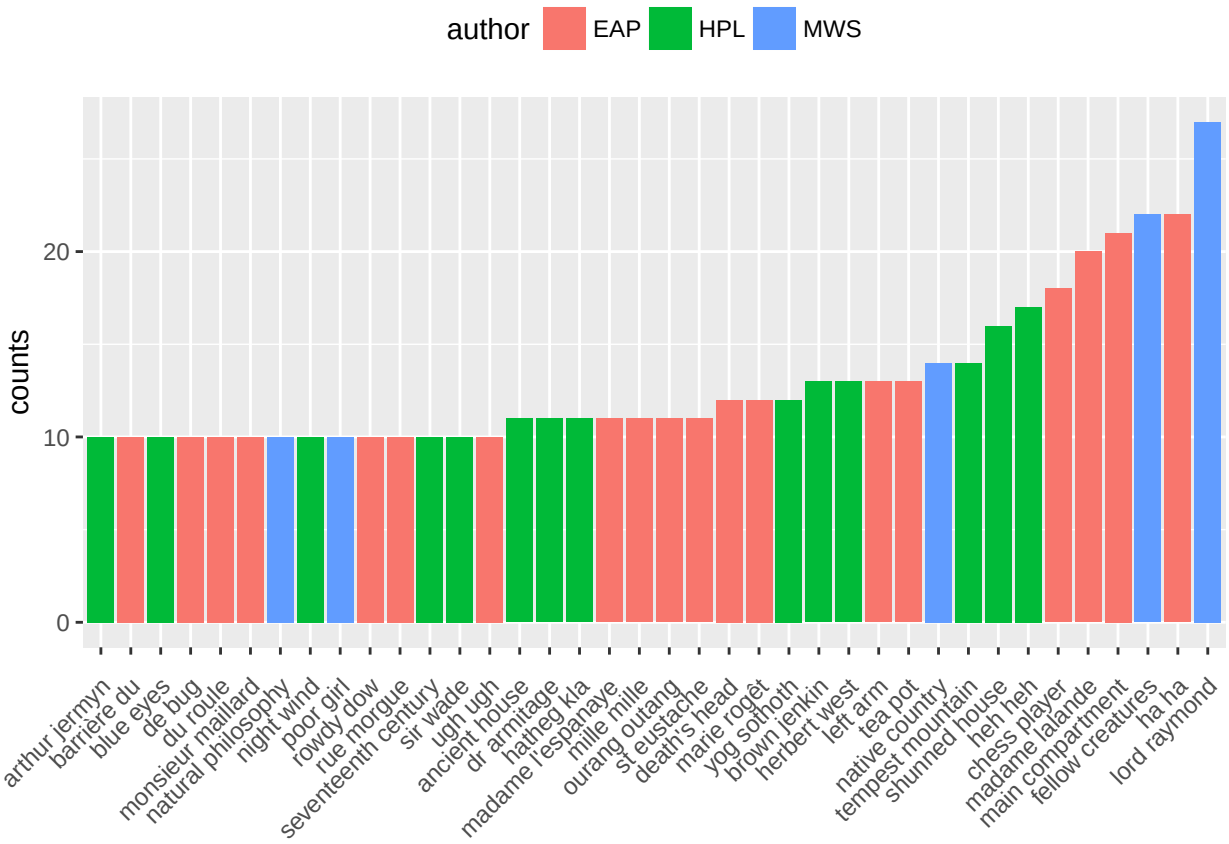
Bigram analysis allows us to observe two consecutive words.

7.1.i Frequency of Bigrams

We create another dataset using the `unnest_tokens()` function but this time to include two words at a time.

We create an increasing ordered bar graph to show the most frequently used bigrams, colored by the author.

```
# Get the bigrams
spooky_wrd_bigrams <- unnest_tokens(spooky, bigram, text, token = "ngrams", n = 2)

# Separate the bigrams into its own column
spooky_wrd_bigrams <- separate(spooky_wrd_bigrams, bigram, c("word1", "word2"), sep = " ")

# Filter out the stop words
spooky_wrd_bigrams <- spooky_wrd_bigrams %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

#Unite the bigrams again
spooky_wrd_bigrams <- unite(spooky_wrd_bigrams, bigram, word1, word2, sep = " ")

#Count the number of unique bigrams for each author
bigram_counts <- count(spooky_wrd_bigrams, author, bigram, sort = T)

# This gives us the top 30 counts
bigram_counts <- top_n(bigram_counts, 30, n)

ggplot(bigram_counts) +
  geom_col(aes(reorder(bigram, n), n, fill = author)) +
  labs(x = NULL, y = "counts") +
  theme(legend.position = "top", axis.text.x  = element_text(angle=45, hjust=1, vjust=0.9))
```



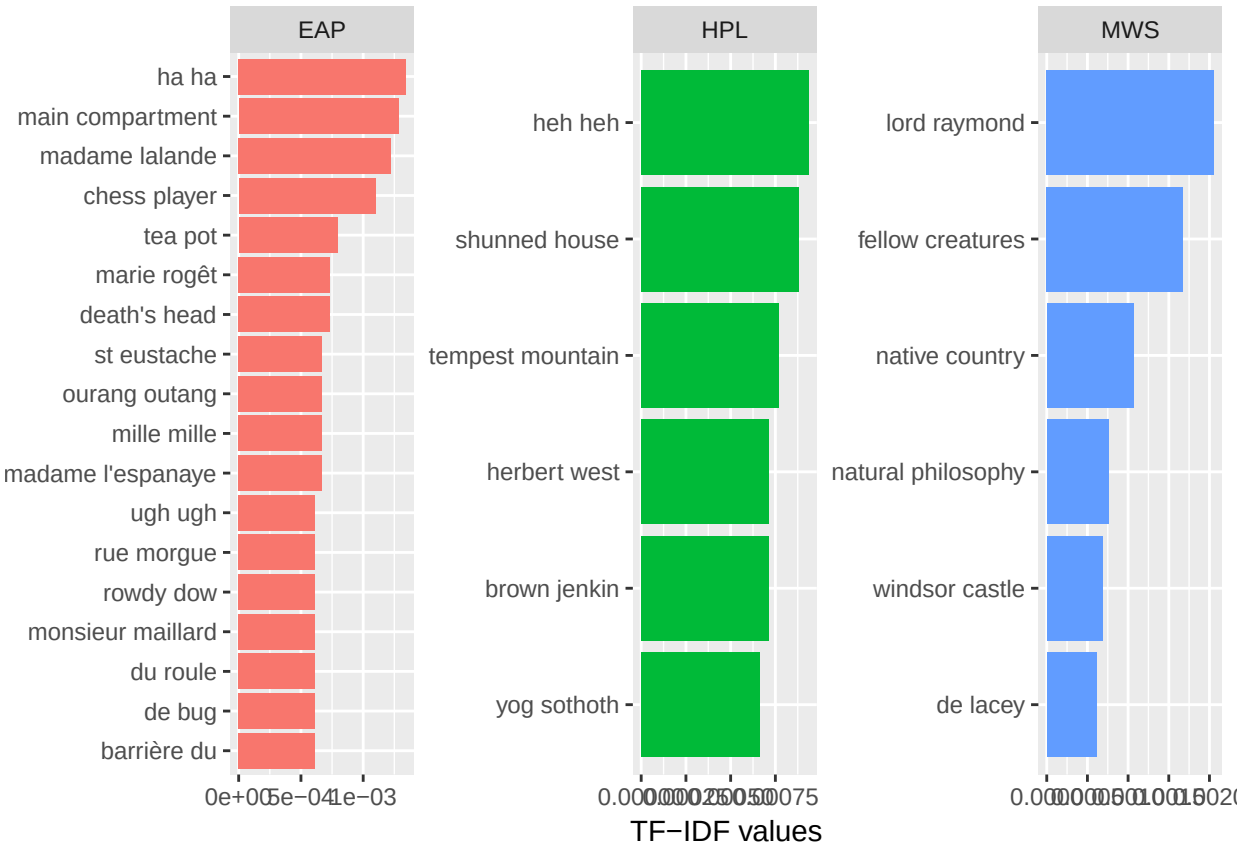
Observations:

1. MWS' 'Lord Raymond' is the most commonly used bigram among all three authors. 'Ha ha' is used most commonly by EAP. 'Heh heh' is used most frequently by HPL.
2. Among the top 30 bigrams, EAP holds the majority of them, whereas MWS holds the least.

7.1.ii Tf_idf

Just for fault-proof, a Tf_idf analysis is also done to look at the top 30 rarest bigrams among the three authors.

```
# Calculate the tf-idf for each bigram
bigram_tf_idf <- spooky_wrd_bigrams %>%
  count(author, bigram) %>%
  bind_tf_idf(bigram, author, n) %>%
  arrange(desc(tf_idf))

# Give the top 30 tf-idfs
bigram_tf_idf <- head(bigram_tf_idf, 30)

ggplot(bigram_tf_idf) +
  geom_col(aes(reorder(bigram, tf_idf), tf_idf, 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")
```



Observations:

1. Some special bigrams for EAP I observe are 'main compartment', 'chess player', 'death's head', and 'rowdy row'.
2. HPL's 'Shunned house', 'tempest mountain', and 'yog sothoth' are also some unique combinations.
3. MWS's 'fellow creatures', 'native country', and 'natural philosophy' are quite frequent occurences.

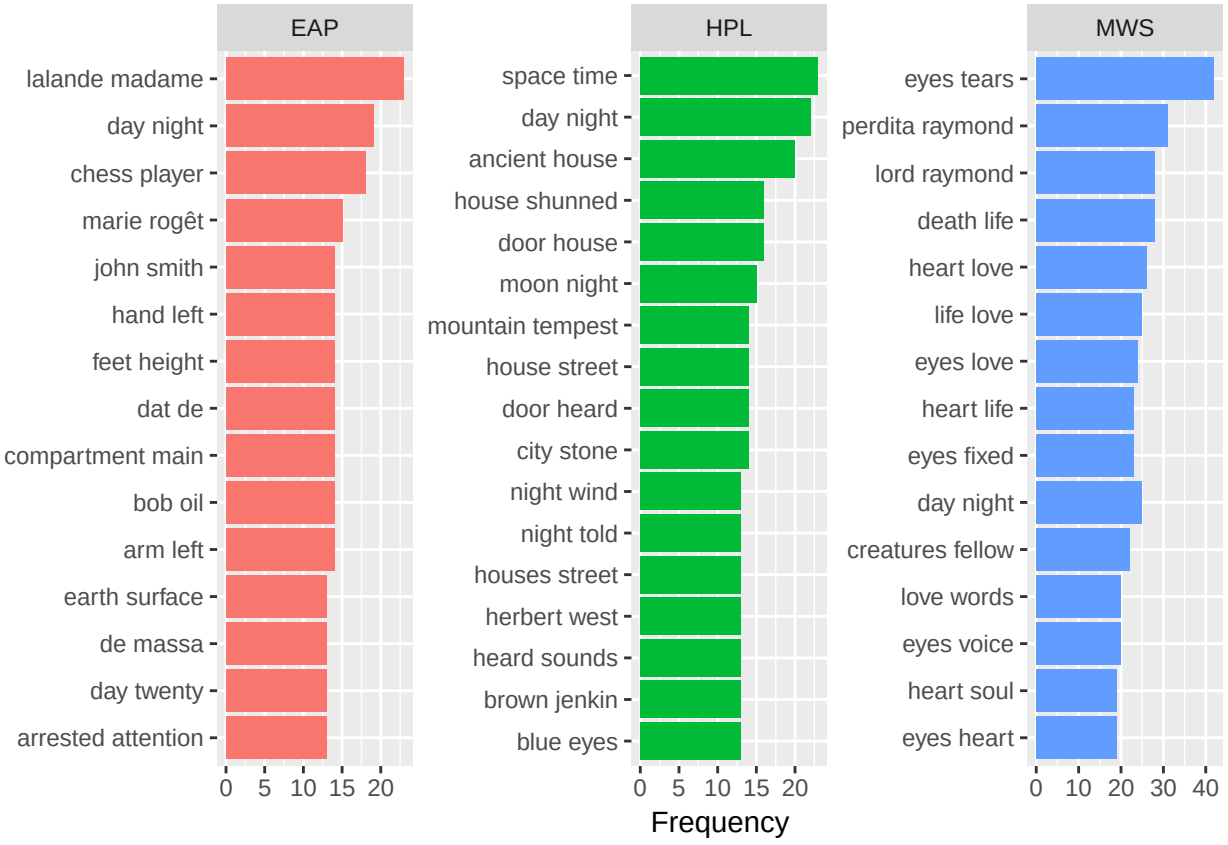
7.2 Word Pairs Dataset

Now we want to analyze any pair of words that might show up in each text. This is separate from our bigram analysis because the pairs are irrespective of their adjacency, as long as they exist in the same text/passage.

7.2.i Frequency of Word Pairs

```
EAP_pairs <- pairwise_count(filter(spooky_wrd, author == "EAP"), word, id, sort = TRUE)
EAP_pairs$author <- "EAP"
HPL_pairs <- pairwise_count(filter(spooky_wrd, author == "HPL"), word, id, sort = TRUE)
HPL_pairs$author <- "HPL"
MWS_pairs <- pairwise_count(filter(spooky_wrd, author == "MWS"), word, id, sort = TRUE)
MWS_pairs$author <- "MWS"

# ddply is used to count pairwise for each author group
pairs <- ddply(spooky_wrd, .(author), pairwise_count, word, id, sort = TRUE)
pairs <- group_by(pairs, author)
```

```r
# Filtered top 30 for faster computation
pairs <- top_n(pairs, 30, n)

# Remove duplicates by sorting
sorted_pairs <- t(apply(pairs[,2:3], 1, sort))
pairs$item1 <- sorted_pairs[, 1]
pairs$item2 <- sorted_pairs[, 2]
pairs <- pairs[!duplicated(pairs),]

# United Pairs into one column
pairs <- unite(pairs, pairs, item1, item2, sep = " ")

ggplot(pairs) +
  geom_col(aes(reorder(pairs, n), n, fill = author)) +
  labs(x = NULL, y = "Frequency") +
  theme(legend.position = "none") +
  facet_wrap(~ author, ncol = 3, scales = "free") +
  coord_flip() +
  labs(y = "Frequency")
```



Observations:

1. EAP's most common pair is 'lalande madame'. Interestingly, 'Chess player' is also a pretty common pair.
2. HPL's most common pair is 'space time', also followed by 'day night'. A lot of the pairs also include words such as 'house', 'door', street'.
3. MWS' most common pair is 'eyes tear'. A lot of her pairs include words such as 'love' and 'eyes'.

4. All three authors have 'day night' in their top pairs. EAP and HPL as their 2nd highest, and MWS as ranked 6th. However it is important to note that the frequency of MWS' pairs are higher than the other two authors. Therefore, she actually uses 'day night' the most.

8 Sentiment Analysis

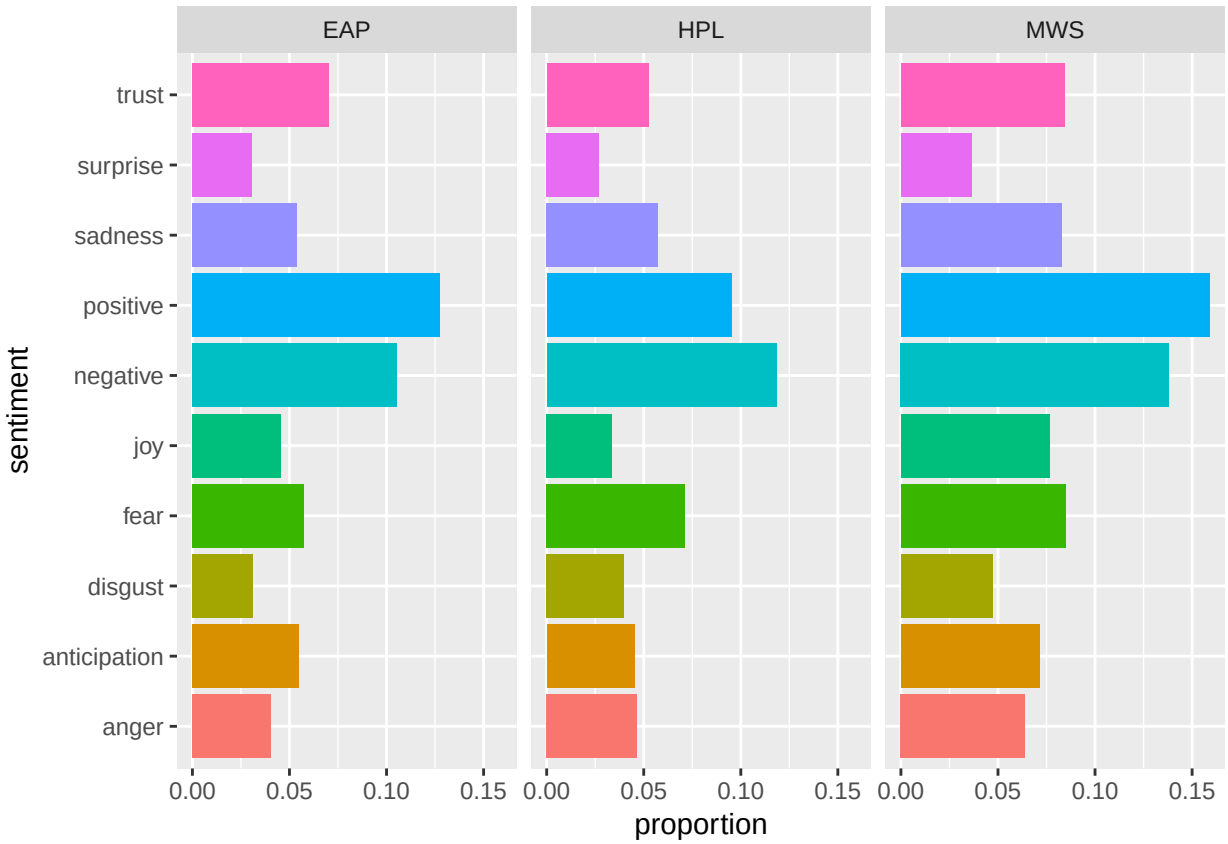
Last but not least, we may want to look at some sentiment analysis to see what kind of emotions are the authors trying to evoke out of the readers.

8.1 Sentiment Words across Authors

```
# Keep words that have been classified within the NRC lexicon.
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
```

```
sentiments <- inner_join(spooky_wrd, get_sentiments('nrc'), by = "word")

# Count how many times each sentiment occurs
sentiments_count <- count(group_by(sentiments, sentiment, author))

# Convert to proportion because each author are featured at different amounts
sentiments_count$proportion <- ifelse(sentiments_count$author == "EAP", sentiments_count$n/frequency$EAP

ggplot(sentiments_count) +
  geom_col(aes(sentiment, proportion, fill = sentiment)) +
  facet_wrap(~ author) +
  coord_flip() +
  theme(legend.position = "none")
```



Observations:

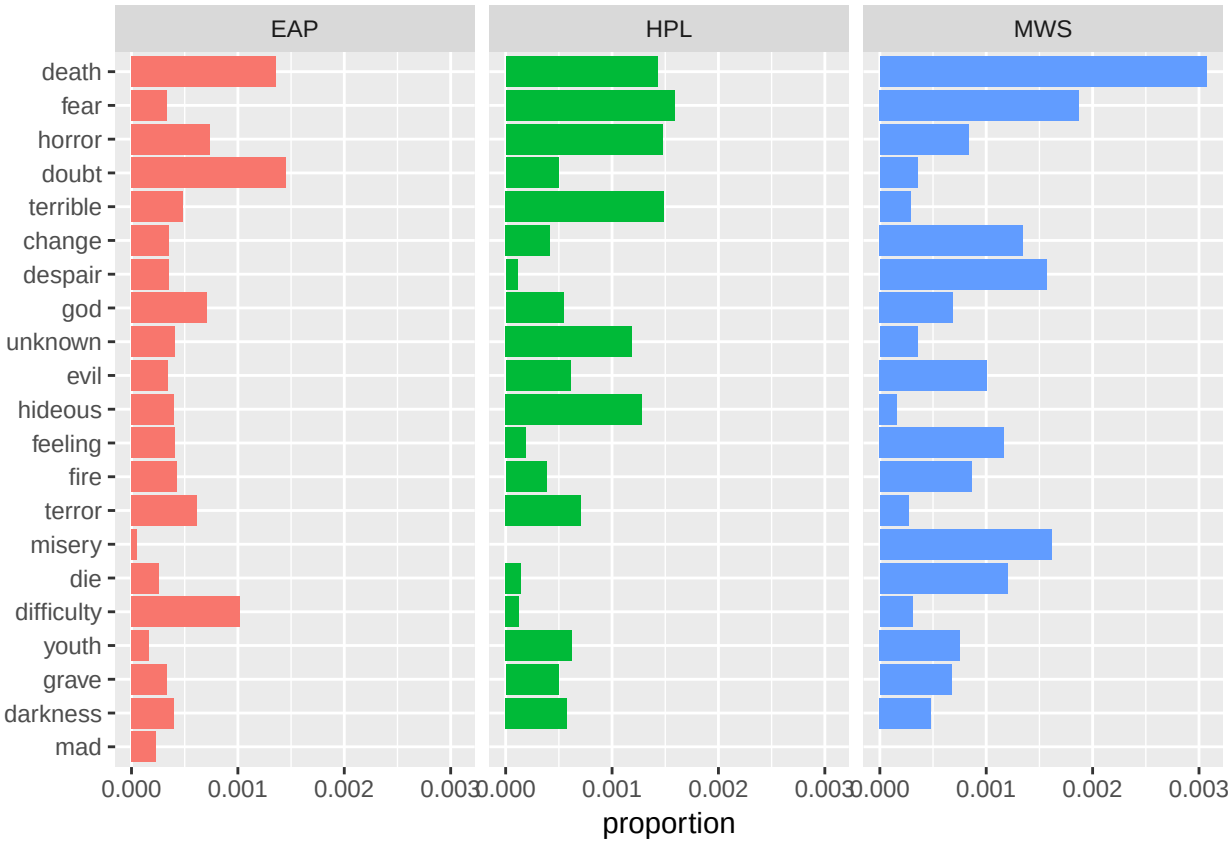
1. EAP does not use many words with fear and disgust, compared to the other authors.
2. EAP has the highest proportion of trust
3. HPL is the only author with more negative words than positive words.
4. HPL has the smallest proportion of joy and surprise.
5. MWS has the highest proportion of positive words, as well as negative words and anticipation.
6. MWS has the highest proportion of words associated with sentiment.

8.2 Fear Dataset

Since we are analyzing horror texts, I thought that looking at the fear sentiment might yield some interest results.

```
nrc_fear <- filter(get_sentiments('nrc'), sentiment == "fear")

fear <- inner_join(spooky_wrd, nrc_fear, by = "word")

#Counts the number of fear words for each author, as well as the total among all three
fear_words     <- count(group_by(fear, word, author))
fear_words_all <- rename(count(group_by(fear, word)), all = n)

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

# Counts the proportion of the word in the total number of words featured by each author
```

```
fear_words$proportion <- ifelse(fear_words$author == "EAP", fear_words$n/frequency$EAP, ifelse(fear_wor

ggplot(fear_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. 'Death' is the most frequent fear word, and MWS has the largest proportion of that word used among all words in MWS text.
2. EAP has the smallest proportion of the word 'fear'. EAP has the largest proportion of 'doubt'
3. 'Terrible' is a popular word for HPL texts, and less for the other two authors.
4. 'Misery' is a word that is used often by MWS, used drastically less by EAP, and not used at all by HPL.
5. 'Horror' is a relatively common word among all three authors.