STAT GU4243 Project 1

Juho Ma (jm4382)

February 5, 2018

Introduction

This project studies the `spooky.csv` data, which contains text excerpts from three popular horror authors, Edgar Allan Poe, HP Lovecraft, and Mary Shelley. By using different methods of text analysis, it attempts to demonstrate similarities and differences among the three authors' texts, and to find patterns that could characterize the writing styles of the authors.

Setting Up the Libraries

```
packages.used <- c("ggplot2", "dplyr", "tibble", "tidyr",  "stringr",
                   "tidytext", "topicmodels", "wordcloud", "ggridges", "forcats")
packages.needed <- setdiff(packages.used, intersect(installed.packages()[,1], packages.used))
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(forcats)
source("../libs/multiplot.R")
```

Read in the Data

The original `spooky.csv` data comes in the format of individual sentences with unique ID's and abbreviated author names. It resides in the `data` folder within the same directory as the `doc` folder's. EAP stands for Edgar Allan Poe, HPL for HP Lovecraft, and MWS for Mary Shelley. There are 7900 sentences for EAP, 5635 for HPL, and 6044 for MWS.

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

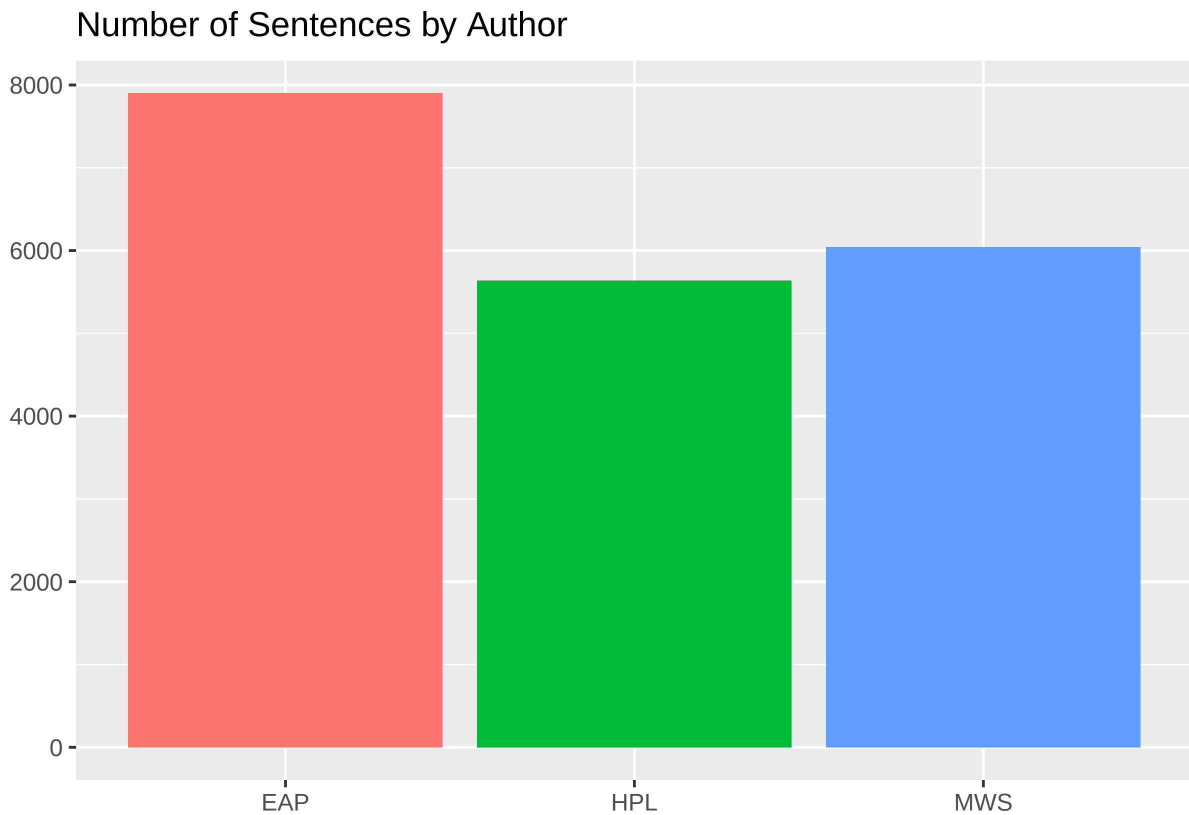
```r
sum(is.na(spooky))
```

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

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

```
##       id                 text            author
##  Length:19579       Length:19579       EAP:7900
##  Class :character   Class :character   HPL:5635
##  Mode  :character   Mode  :character   MWS:6044
```

```r
ggplot(spooky) +
  geom_bar(aes(author, fill = author)) +
  ggtitle("Number of Sentences by Author") +
  ylab("") +
  xlab("") +
  theme(legend.position = "none")
```



Stylistic Analysis

Length of Sentences

Length of sentence is oftentimes one of the most distinctive characteristics of one's writing. We can measure the length of sentences from two different perspectives: 1) the number of characters in a sentence, and 2) the number of words in a sentence.

```
spooky$nchar <- str_length(spooky$text)
wordcount <- function(str) {
  sapply(gregexpr("\\b\\W+\\b", str, perl=TRUE), function(x) sum(x>0) ) + 1
}
spooky$nword <- wordcount(spooky$text)
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 re
##   author nchar nword
## 1    EAP   231    41
## 2    HPL    71    14
## 3    EAP   200    36
## 4    MWS   206    34
## 5    HPL   174    27
## 6    MWS   468    83
```

```
summary_nchar <- spooky %>%
                 group_by(author) %>%
                 summarize(mean = mean(nchar), median = median(nchar), std = sd(nchar))
summary_nword <- spooky %>%
                 group_by(author) %>%
                 summarize(mean = mean(nword), median = median(nword), std = sd(nword))
summary_nchar
```

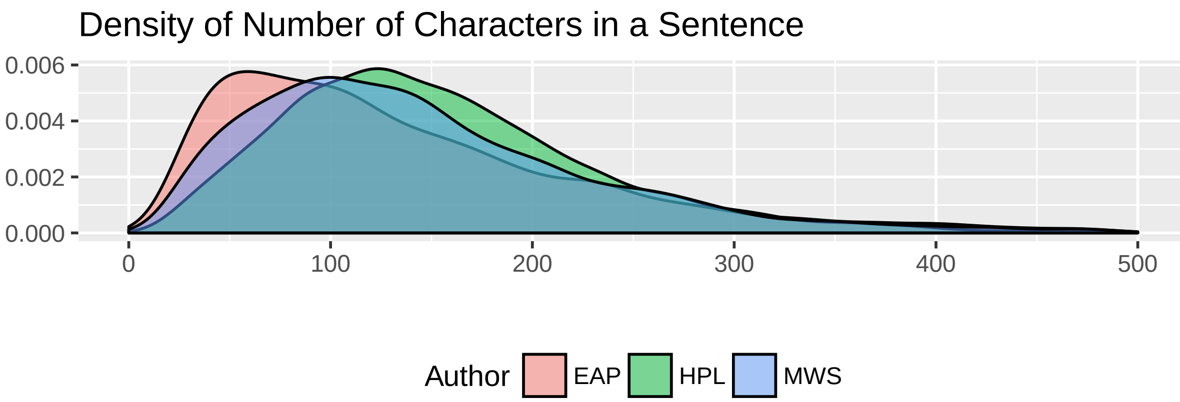
```
## # A tibble: 3 x 4
##   author  mean median   std
##   <fct>  <dbl>  <dbl> <dbl>
## 1 EAP      142    115 106
## 2 HPL      156    142  82.0
## 3 MWS      152    130 126
```

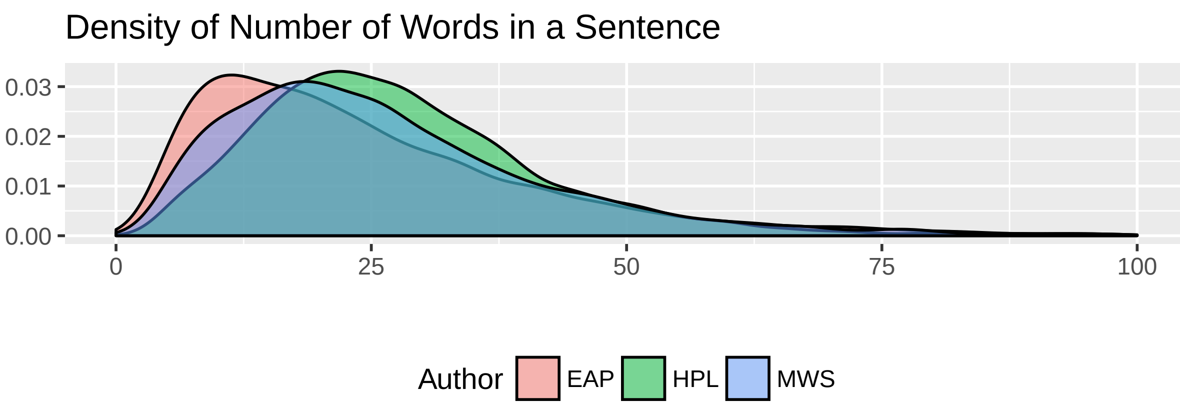
```
summary_nword
```

```
## # A tibble: 3 x 4
##   author  mean median   std
##   <fct>  <dbl>  <dbl> <dbl>
## 1 EAP     25.5   21.0  18.6
## 2 HPL     27.9   26.0  14.2
## 3 MWS     27.5   23.0  23.2
```

```
p1 <- ggplot(spooky, aes(nchar, ..density.., fill = author)) +
        geom_density(alpha = .5) +
        xlim(0, 500) +
        xlab("") +
        ylab("") +
        labs(fill = 'Author') +
        ggtitle("Density of Number of Characters in a Sentence") +
        theme(legend.position = "bottom")

p2 <- ggplot(spooky, aes(nword, ..density.., fill = author)) +
        geom_density(alpha = .5) +
        xlim(0, 100) +
        xlab("") +
        ylab("") +
        labs(fill = 'Author') +
        ggtitle("Density of Number of Words in a Sentence") +
        theme(legend.position = "bottom")

multiplot(p1, p2)
```

```
## Warning: Removed 158 rows containing non-finite values (stat_density).

## Warning: Removed 89 rows containing non-finite values (stat_density).
```





From the summary, we can see that the two measures of the length of sentence, 1) the number of characters in a sentence and 2) the number of words in a sentence show a similar pattern. If you compare this result among three authors, HPL generally uses the longest sentences among three, both in terms of the number of characters and the number of words, and EAP the shortest among three.

Use of Punctuation Marks

The use of punctuation marks in a sentence can also significantly characterize one's writing style. Although previously we could see HPL's sentences are generally longer than the other two, here, we see that he uses less punctuation marks such as commas, semicolons, and colons. Although the differences are minor, we can argue that longer sentences with fewer punctuation marks is a distinctive characteristic of HPL.

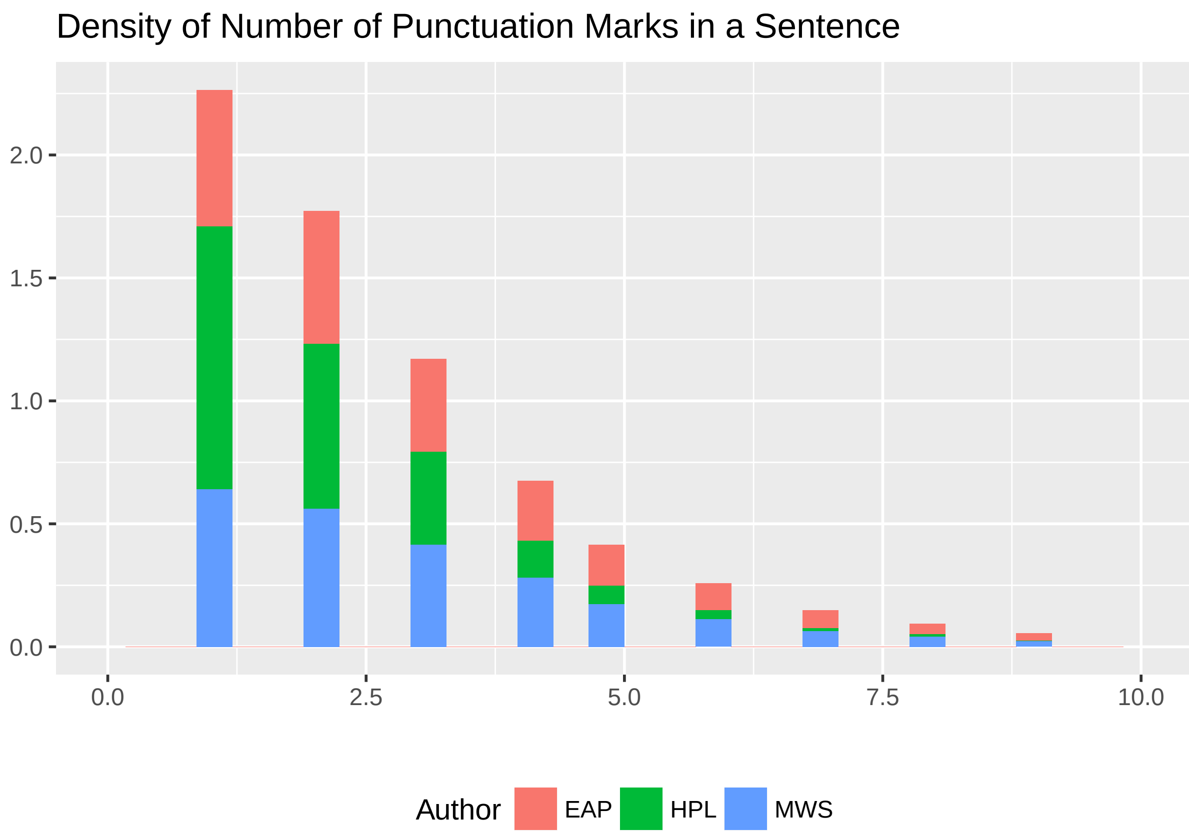
```
spooky$npunct <- str_count(spooky$text, ",") + str_count(spooky$text, ";") + str_count(spooky$text, ":"
summary_npunct <- spooky %>%
                  group_by(author) %>%
                  summarize(mean = mean(npunct), median = median(npunct), std = sd(npunct))
summary_npunct
```

```
## # A tibble: 3 x 4
##   author  mean median   std
##   <fct>  <dbl>  <dbl> <dbl>
## 1 EAP     2.42   2.00  2.63
## 2 HPL     1.73   1.00  1.53
## 3 MWS     2.49   2.00  2.46
```

```
ggplot(spooky, aes(npunct, ..density.., fill = author)) +
  geom_histogram(alpha = 1) +
  xlim(0, 10) +
```

```
  xlab("") +
  ylab("") +
  labs(fill = 'Author') +
  ggtitle("Density of Number of Punctuation Marks in a Sentence") +
  theme(legend.position = "bottom")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 192 rows containing non-finite values (stat_bin).
```



Use of Vocabulary

The type of words one uses can also identify the writer's style. In order to analyze the data on the word level, we are going to create `spooky_word` that lists words instead of sentences.

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

```
##          id author nchar nword npunct     word
## 1   id26305    EAP   231    41      6     this
## 1.1 id26305    EAP   231    41      6  process
## 1.2 id26305    EAP   231    41      6  however
## 1.3 id26305    EAP   231    41      6 afforded
## 1.4 id26305    EAP   231    41      6       me
## 1.5 id26305    EAP   231    41      6       no
```

Length of Words

The first characteristic we can look at is the length of each word the writers use. Unfortunately, we don't see a great difference among three authors.

```
spooky_word$word_nchar <- str_length(spooky_word$word)
head(spooky_word)
```

```
##          id author nchar nword npunct     word word_nchar
## 1   id26305    EAP   231    41      6     this          4
## 1.1 id26305    EAP   231    41      6  process          7
## 1.2 id26305    EAP   231    41      6  however          7
## 1.3 id26305    EAP   231    41      6 afforded          8
## 1.4 id26305    EAP   231    41      6       me          2
## 1.5 id26305    EAP   231    41      6       no          2
```

```
summary_word_nchar <- spooky_word %>%
                      group_by(author) %>%
                      summarize(mean = mean(word_nchar), median = median(word_nchar), std = sd(word_ncha
summary_word_nchar
```

```
## # A tibble: 3 x 4
##   author  mean median   std
##   <fct>  <dbl>  <dbl> <dbl>
## 1 EAP     4.47   4.00  2.53
## 2 HPL     4.54   4.00  2.40
## 3 MWS     4.43   4.00  2.40
```

The result is pretty much the same after eliminating stop words from the data. The graph shows similar distributions among three authors, although EAP's shows slightly greater use of longer words.

```
spooky_word <- anti_join(spooky_word, stop_words, by = "word")
spooky_word$word_nchar <- str_length(spooky_word$word)
head(spooky_word)
```

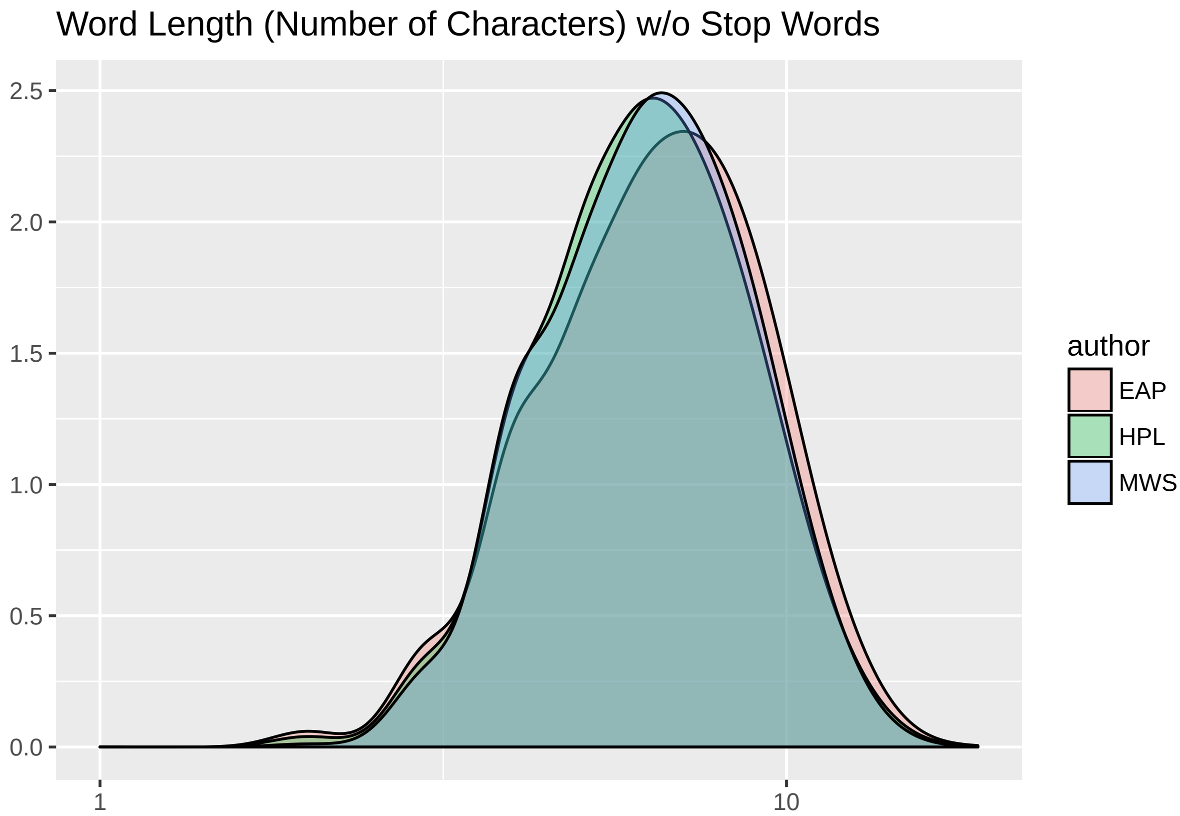
```
##        id author nchar nword npunct         word word_nchar
## 1 id26305    EAP   231    41      6      process          7
## 2 id26305    EAP   231    41      6     afforded          8
## 3 id26305    EAP   231    41      6        means          5
## 4 id26305    EAP   231    41      6 ascertaining         12
## 5 id26305    EAP   231    41      6   dimensions         10
## 6 id26305    EAP   231    41      6      dungeon          7
```

```
summary_word_nchar <- spooky_word %>%
                      group_by(author) %>%
                      summarize(mean = mean(word_nchar), median = median(word_nchar), std = sd(word_ncha
summary_word_nchar
```

```
## # A tibble: 3 x 4
##   author  mean median   std
##   <fct>  <dbl>  <dbl> <dbl>
## 1 EAP     6.79   7.00  2.36
## 2 HPL     6.53   6.00  2.19
## 3 MWS     6.60   6.00  2.15
```

```
ggplot(spooky_word) +
  geom_density(aes(word_nchar, fill = author), bw = 0.05, alpha = 0.3) +
  scale_x_log10() +
```

```
  xlab("") +
  ylab("") +
  ggtitle("Word Length (Number of Characters) w/o Stop Words")
```

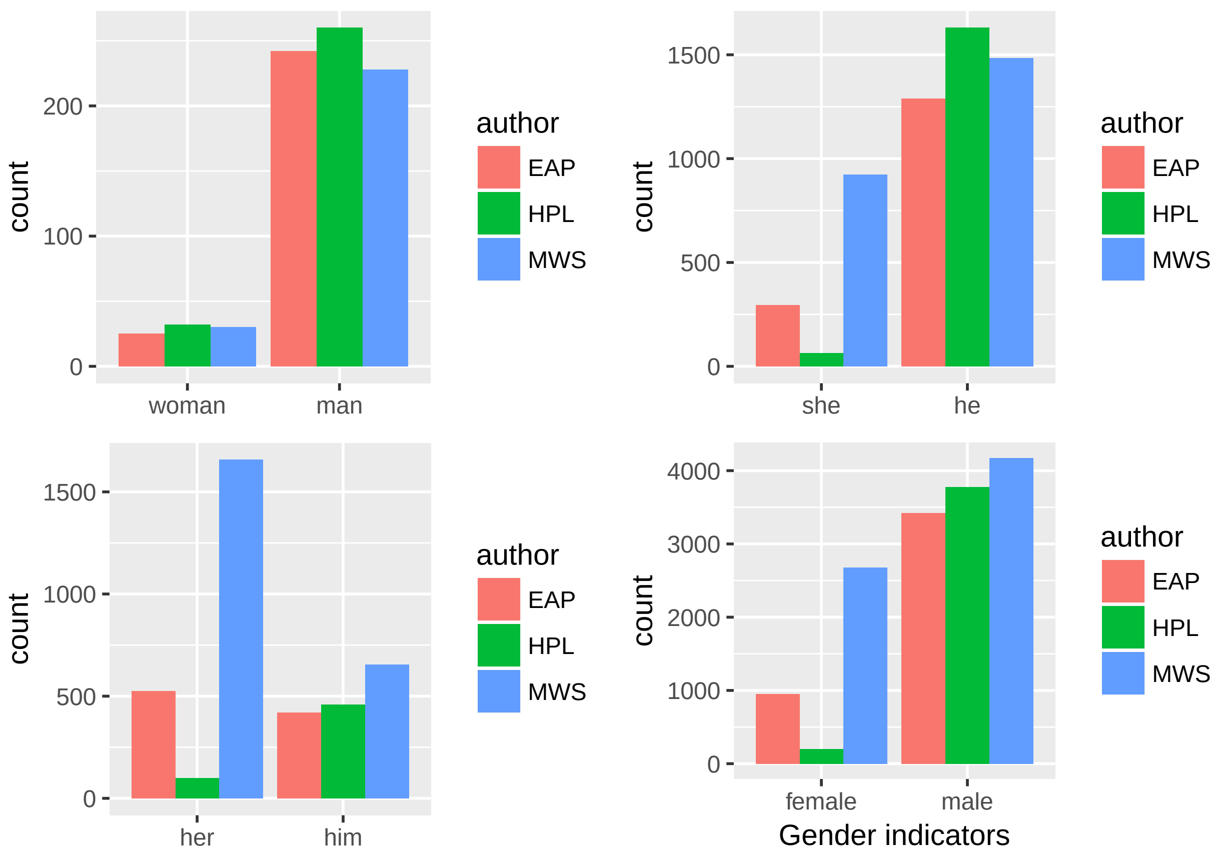


Gender Representation

We can also look at how different genders are represented in each author's writings. By comparing `man` and `woman`, `he` and `she`, `him` and `her`, and other male and female gender indicators, we will see how three authors differ in their preference in the gender representation.

```
p1 <- spooky %>%
  unnest_tokens(word, text) %>%
  filter((word == "man") | (word == "woman")) %>%
  mutate(word = as.factor(word)) %>%
  mutate(word = fct_relevel(word, "woman", "man")) %>%
  ggplot(aes(word, fill = author)) +
  geom_bar(position = "dodge") +
  theme(axis.title.x=element_blank())

p2 <- spooky %>%
  unnest_tokens(word, text) %>%
  filter((word == "he") | (word == "she")) %>%
  mutate(word = as.factor(word)) %>%
  mutate(word = fct_relevel(word, "she", "he")) %>%
  ggplot(aes(word, fill = author)) +
  geom_bar(position = "dodge") +
  theme(axis.title.x=element_blank())
```

```
p3 <- spooky %>%
  unnest_tokens(word, text) %>%
  filter((word == "him") | (word == "her")) %>%
  ggplot(aes(word, fill = author)) +
  geom_bar(position = "dodge") +
  theme(axis.title.x=element_blank())

p4 <- spooky %>%
  unnest_tokens(word, text) %>%
  mutate(male = ( word == "he" | word == "him" | word == "his" | word == "male" |
                  word == "man" | word == "gentleman" | word == "sir" |
                  word == "lord" | word == "men" )) %>%
  mutate(female = ( word == "she" | word == "her" | word == "hers" | word == "female" |
                  word == "woman" | word == "lady" | word == "madam" |
                  word == "women" )) %>%
  unite(sex, male, female) %>%
  mutate(sex = fct_recode(as.factor(sex), male = "TRUE_FALSE",
                          female = "FALSE_TRUE", other = "FALSE_FALSE")) %>%
  filter(sex != "other") %>%
  ggplot(aes(sex, fill = author)) +
  labs(x = "Gender indicators") +
  geom_bar(position = "dodge")

layout <- matrix(c(1,2,3,4),2,2,byrow=TRUE)
multiplot(p1, p2, p3, p4, layout=layout)
```



Overall, HPL rarely uses female representation in his writing. MWS, on the other hand, dominantly uses female indicators in her writing, which may not be so surprising considering she herself was a female writer. The difference between EAP and HPL in use of `she`, `her`, or other female indicators is also significant, which means it can be a distinguishing characteristic between the two.

Word Frequency

Looking at the entire dataset, we can see that words like `time`, `eyes`, `life`, `night`, `day`, and `found` are most frequently used among three authors.

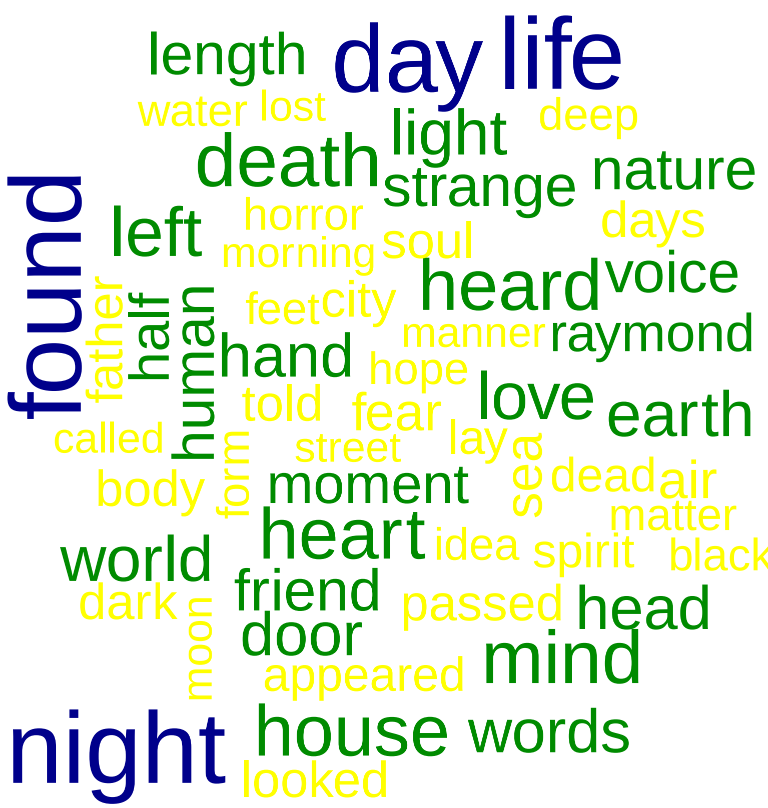
```r
words <- count(group_by(spooky_word, word))$word
freqs <- count(group_by(spooky_word, word))$n

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

```
## [1] 729 563 559 559 540 516
```

```r
wordcloud(words, freqs, max.words = 60, color = c("yellow", "green4", "blue4"))
```

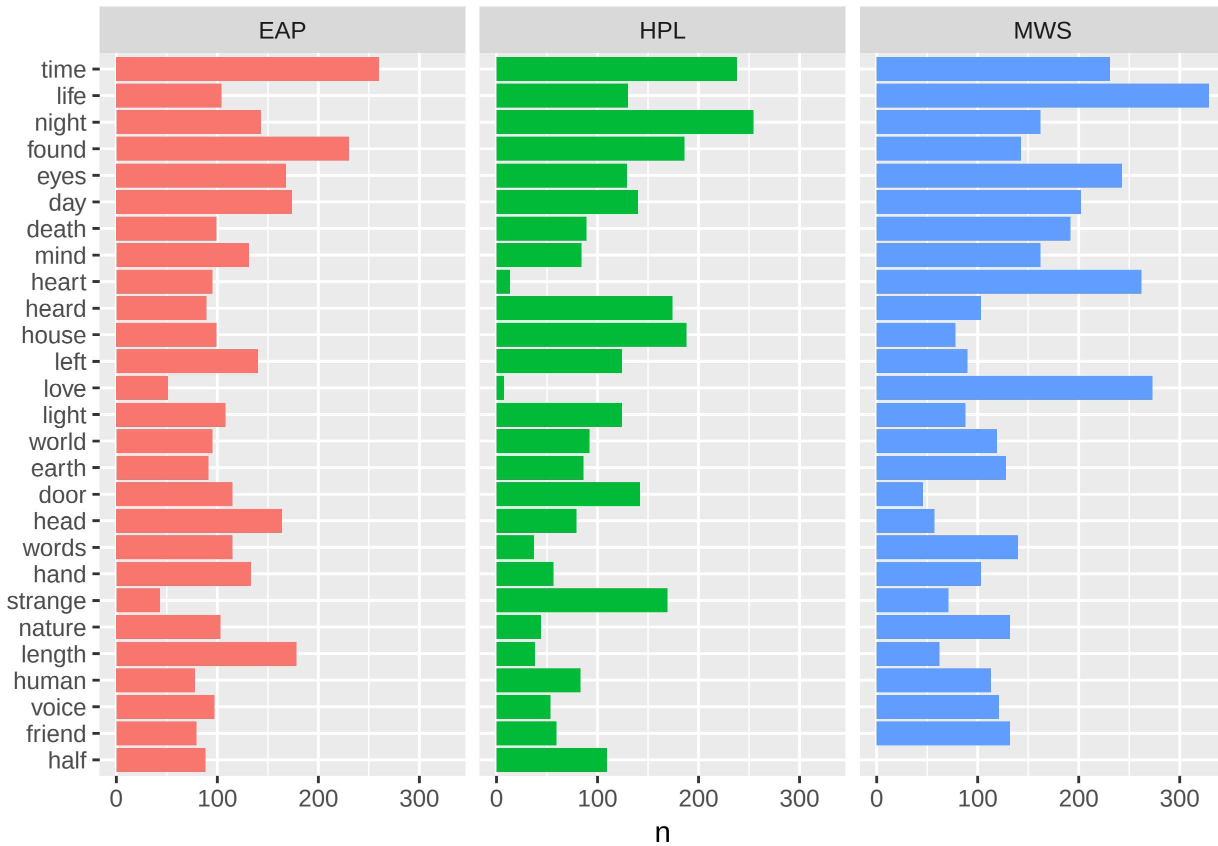
```
## Warning in wordcloud(words, freqs, max.words = 60, color = c("yellow",
## "green4", : time could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words, freqs, max.words = 60, color = c("yellow",
## "green4", : eyes could not be fit on page. It will not be plotted.
```




```r
author_words <- count(group_by(spooky_word, word, author))
all_words <- rename(count(group_by(spooky_word, 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, 80))
```

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



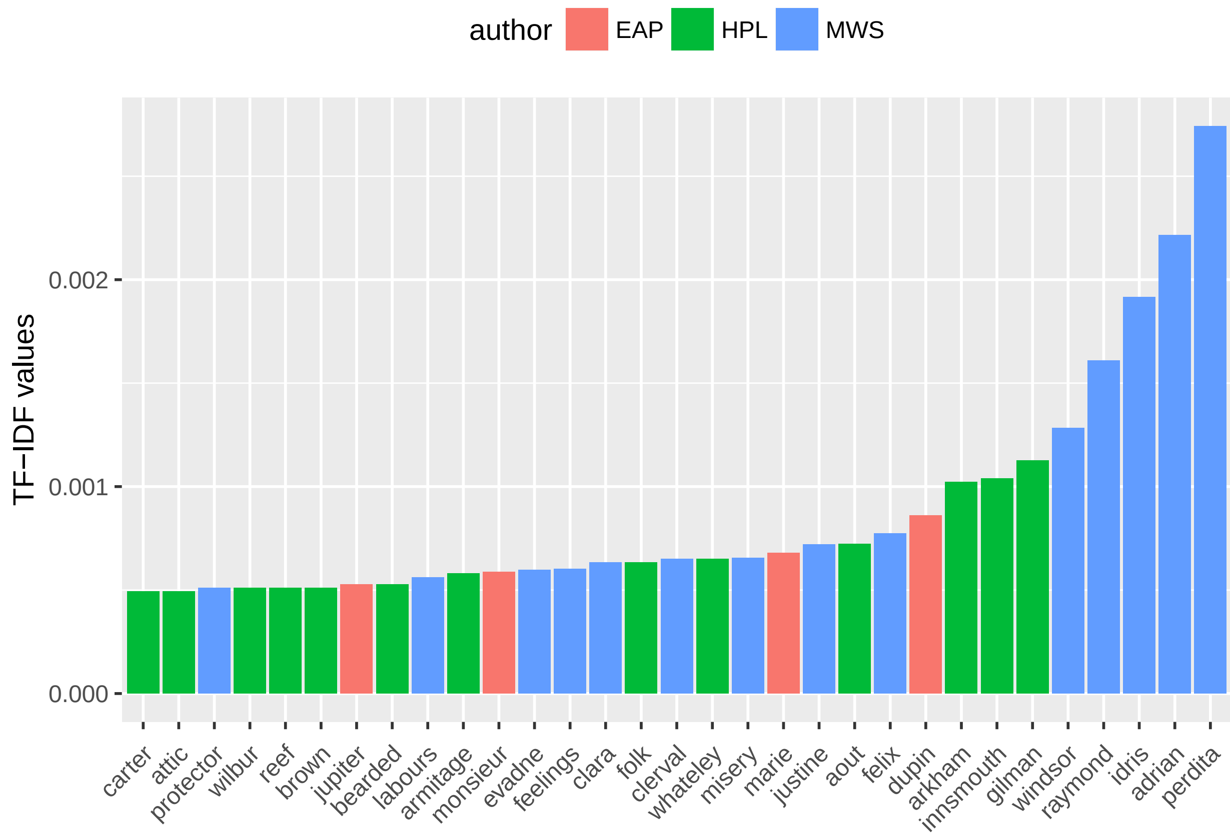
Among the most frequently used words, some words are preferred by specific authors and some not. For example, MWS uses words like `life`, `love`, and `heart` a lot more frequently compared to the other two authors.

TF-IDF

TF-IDF is a heuristic index that indicates how frequent a word is used by a certain author relative to other authors. From this analysis, we can find words that are characteristic for a specific author, which can be useful in identifying the author from a given text.

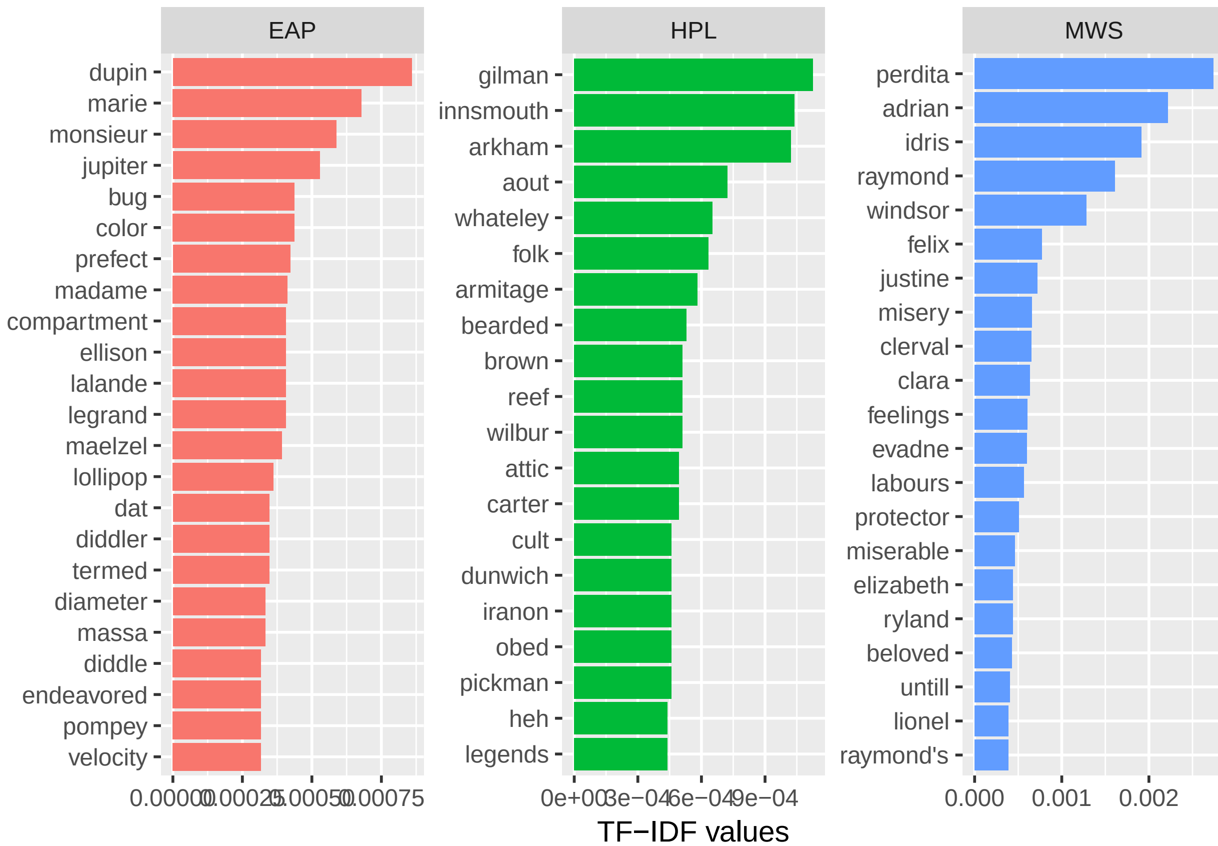
```r
frequency <- count(spooky_word, author, word)
tf_idf <- bind_tf_idf(frequency, word, author, n)
tf_idf <- arrange(tf_idf, desc(tf_idf))
tf_idf <- mutate(tf_idf, word = factor(word, levels = rev(unique(word))))

tf_idf_30 <- top_n(tf_idf, 30, tf_idf)

ggplot(tf_idf_30) +
  geom_col(aes(word, tf_idf, fill = author)) +
```

```r
  labs(x = NULL, y = "TF-IDF values") +
  theme(legend.position = "top", axis.text.x  = element_text(angle=45, hjust=1, vjust=0.9))
```



From looking at the result, we can see that MWS has the most characteristic use of vocabulary of all three authors, especially in her use of names. These names appear a lot more frequently in her writings than other names do in EAP or HPL's writings.

```r
tf_idf <- ungroup(top_n(group_by(tf_idf, author), 20, tf_idf))

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



Among the non-name words, HPL's use of words such as `folk`, `bearded`, `brown`, or `attic` stands out, as they share similar associations that may be characteristic to the settings in his writings. Words such as `monsieur` or `madame` in EAP's writing can be an indication that the settings in his stories are quite different from those in the other two authors' stories.

Bigrams

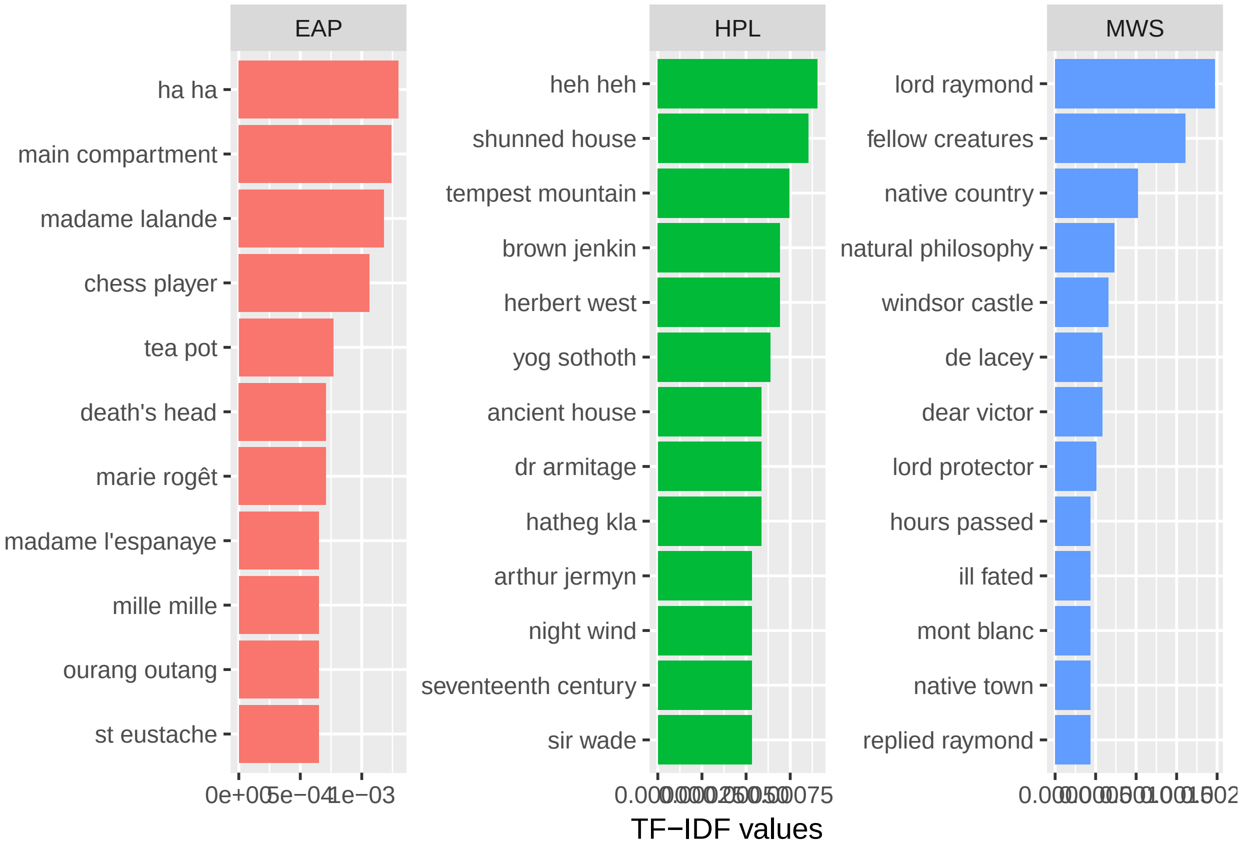
How words appear together can also be an important indicator for the author's identity. We can look at bigrams, a pair of two words, and find meaningful patterns in each author's writings. After filtering out stop words, the TF_IDF values for each author's bigrams appear as follows:

```
t2 <- spooky %>% select(author, text) %>% unnest_tokens(bigram, text, token = "ngrams", n = 2)
sample_n(t2, 5)
```

```
##        author        bigram
## 42688     EAP withdraw his
## 36547     EAP       in fact
## 226390    HPL     the april
## 352845    HPL       for his
## 193218    EAP       a young
```

```
bi_sep <- t2 %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bi_filt <- bi_sep %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)
```

```r
bigram_counts <- bi_filt %>%
  count(word1, word2, sort = TRUE)

t2 <- bi_filt %>%
  unite(bigram, word1, word2, sep = " ")

t2_tf_idf <- t2 %>%
  count(author, bigram) %>%
  bind_tf_idf(bigram, author, n) %>%
  arrange(desc(tf_idf))

t2_tf_idf %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigram = factor(bigram, levels = rev(unique(bigram)))) %>%
  group_by(author) %>%
  top_n(10, tf_idf) %>%
  ungroup() %>%
  ggplot(aes(bigram, tf_idf, fill = author)) +
  geom_col() +
  labs(x = NULL, y = "TF-IDF values") +
  theme(legend.position = "none") +
  facet_wrap(~ author, ncol = 3, scales = "free") +
  coord_flip()
```



Here, we can also see EAP's characteristic settings from words like `main compartment`, `chess player`, or `tea pot`. HPL's more natural settings, as previously indicated before, can also be seen from words such as `tempest mountain` or `night wind`.

Sentiment Analysis

We can analyze sentiments of a sentence by looking at the words and its positive/negative sentiments in it. Here, we apply NRC sentiment lexicon which associates words with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (positive and negative).

```r
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_word, get_sentiments('nrc'), by = "word")

count(sentiments, sentiment)
```

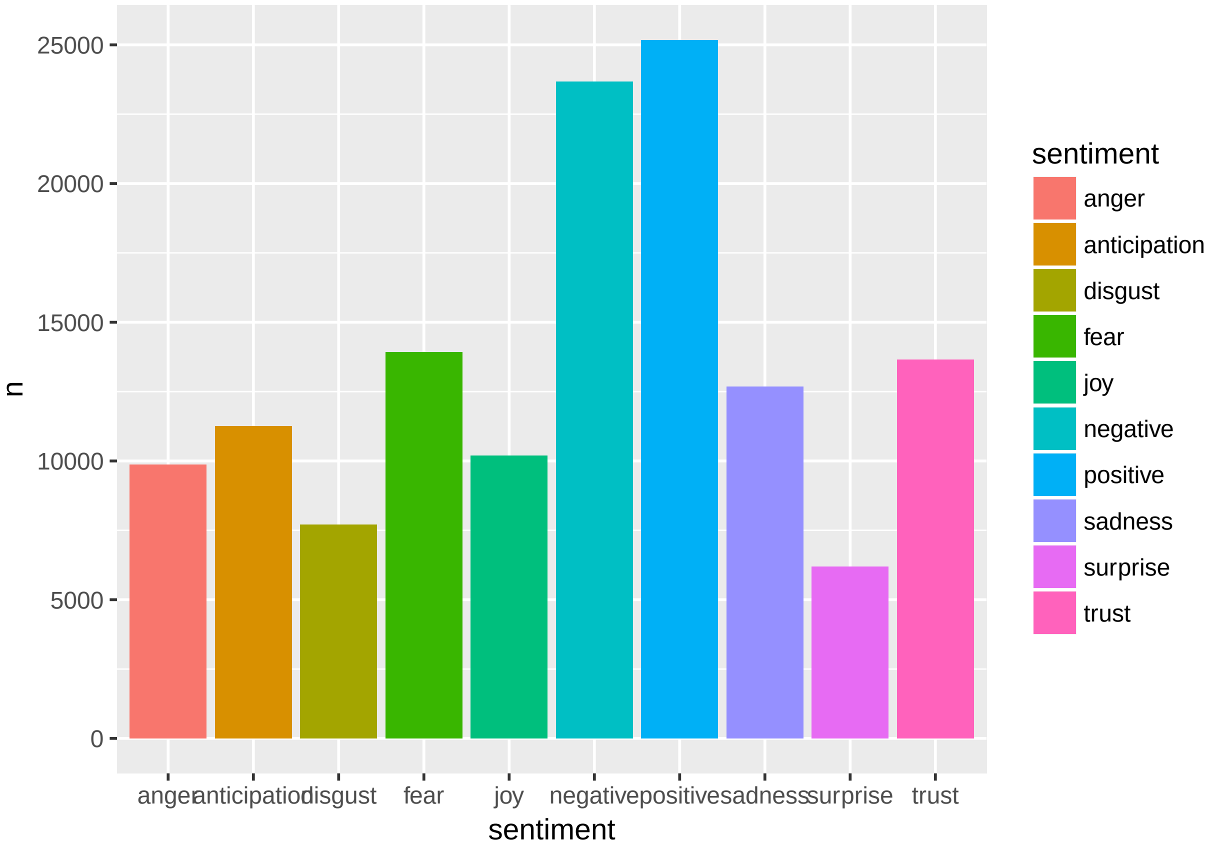
```
## # A tibble: 10 x 2
##    sentiment        n
##    <chr>        <int>
##  1 anger         9869
##  2 anticipation 11258
##  3 disgust       7697
##  4 fear         13927
##  5 joy          10190
##  6 negative     23674
##  7 positive     25175
##  8 sadness      12674
##  9 surprise      6199
## 10 trust        13655
```

```r
count(sentiments, author, sentiment)
```

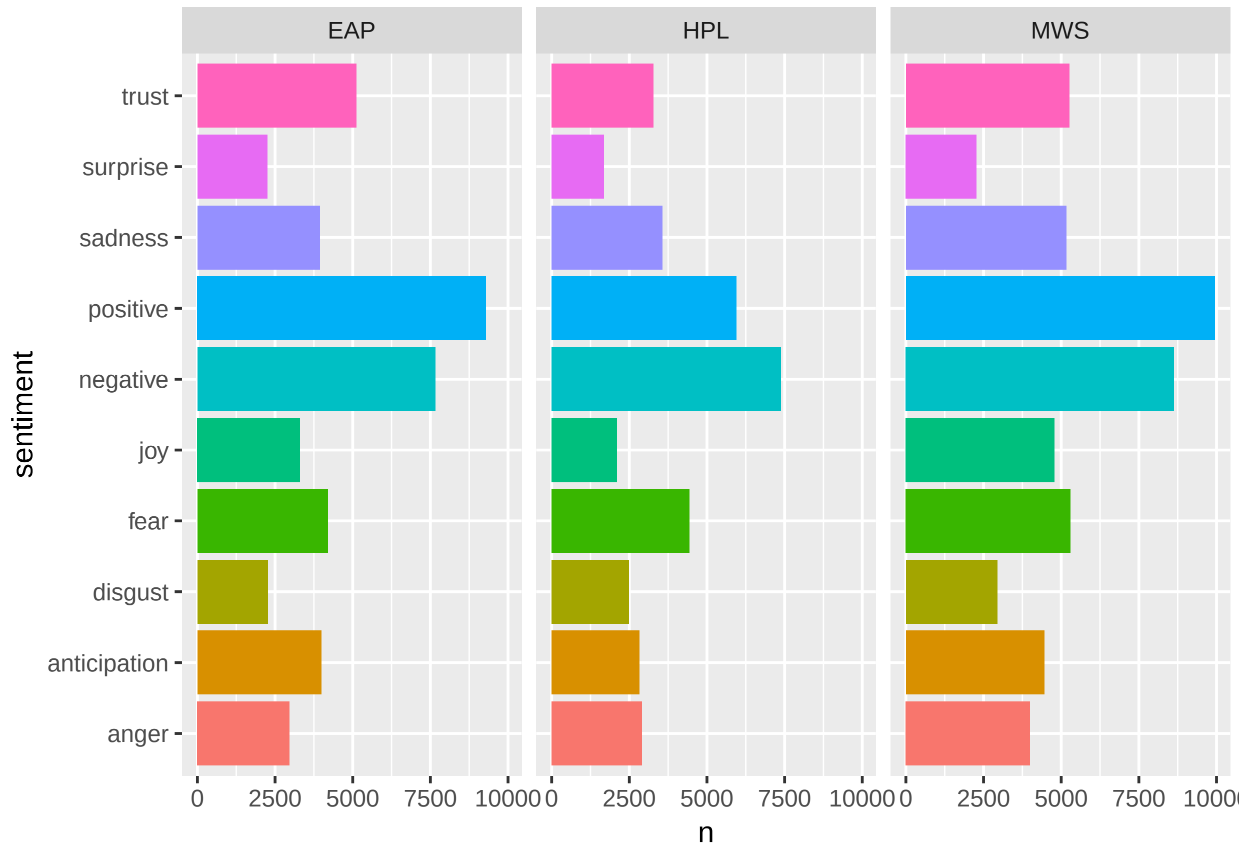
```
## # A tibble: 30 x 3
##    author sentiment        n
##    <fct>  <chr>        <int>
##  1 EAP    anger         2962
##  2 EAP    anticipation  3982
##  3 EAP    disgust       2261
##  4 EAP    fear          4194
##  5 EAP    joy           3302
##  6 EAP    negative      7659
##  7 EAP    positive      9291
```

```
## 8 EAP     sadness       3938
## 9 EAP     surprise      2244
## 10 EAP    trust         5116
## # ... with 20 more rows
```

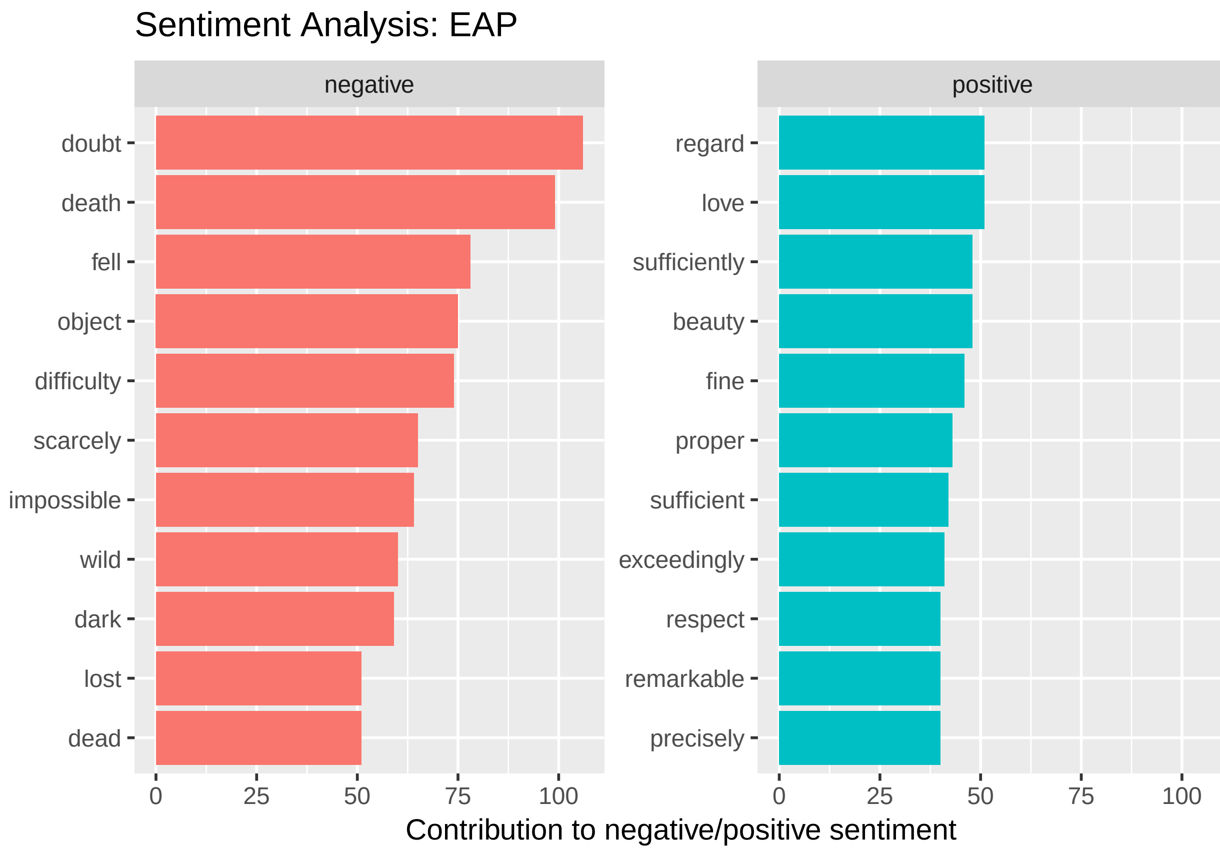
```
ggplot(count(sentiments, sentiment)) +
  geom_col(aes(sentiment, n, fill = sentiment))
```



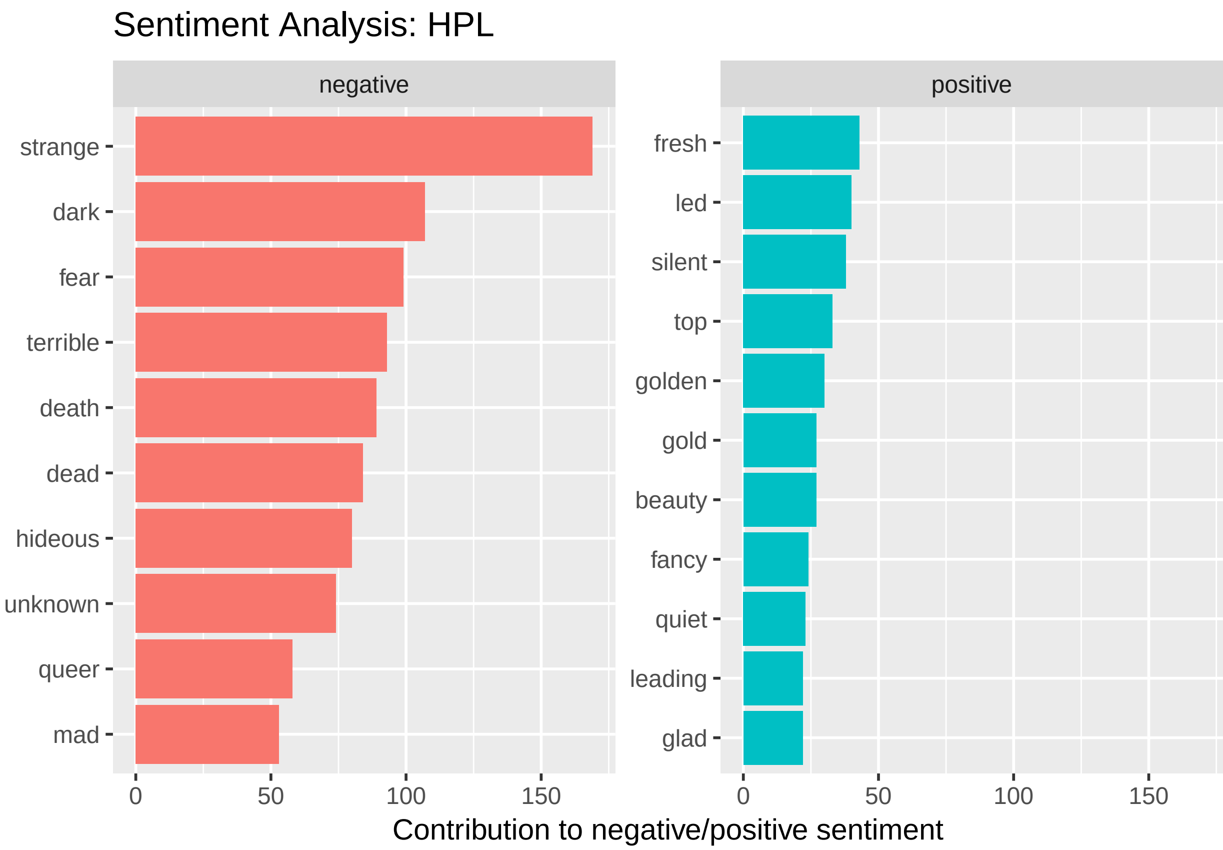

```
ggplot(count(sentiments, author, sentiment)) +
  geom_col(aes(sentiment, n, fill = sentiment)) +
  facet_wrap(~ author) +
  coord_flip() +
  theme(legend.position = "none")
```



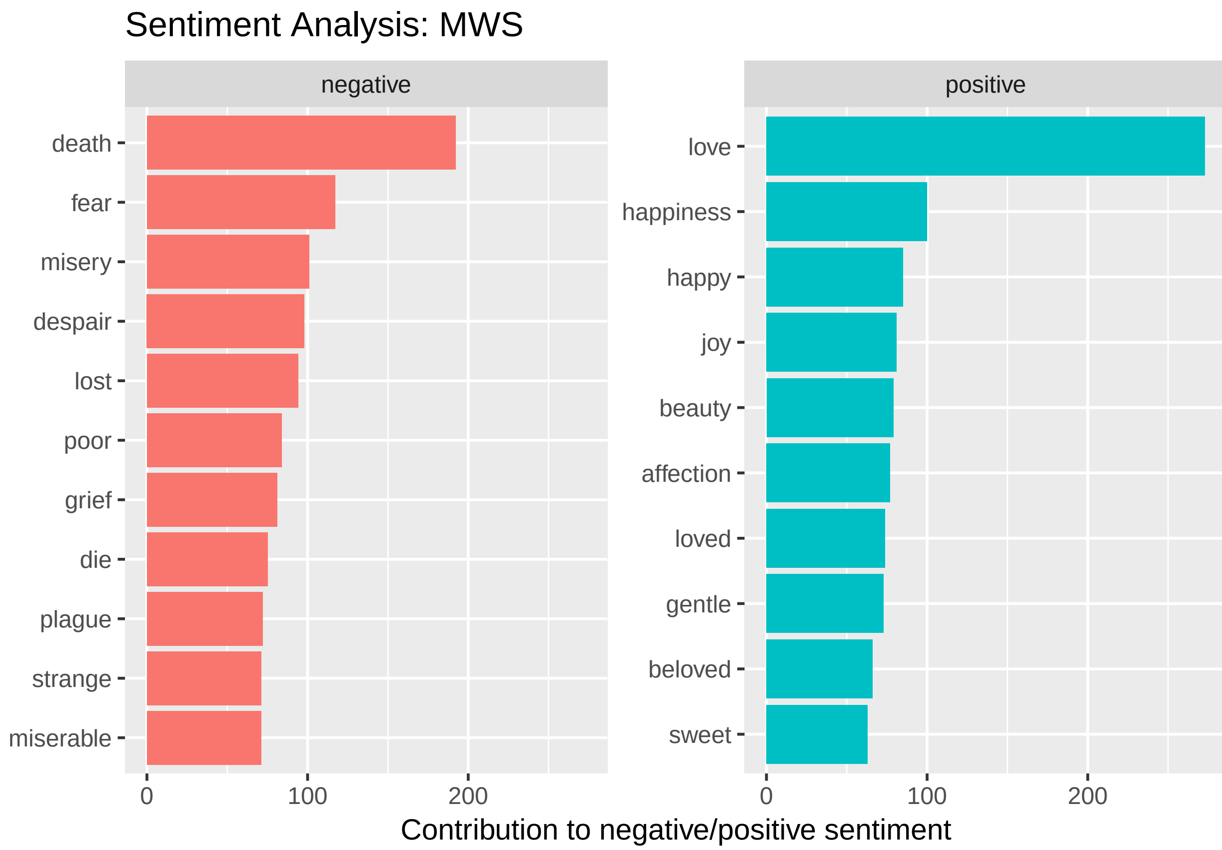

```
spooky_word %>%
  filter(author == "EAP") %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup() %>%
  group_by(sentiment) %>%
  top_n(10, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to negative/positive sentiment", x = NULL) +
  coord_flip() +
  ggtitle("Sentiment Analysis: EAP")
```




```
spooky_word %>%
  filter(author == "HPL") %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup() %>%
  group_by(sentiment) %>%
  top_n(10, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to negative/positive sentiment", x = NULL) +
  coord_flip() +
  ggtitle("Sentiment Analysis: HPL")
```



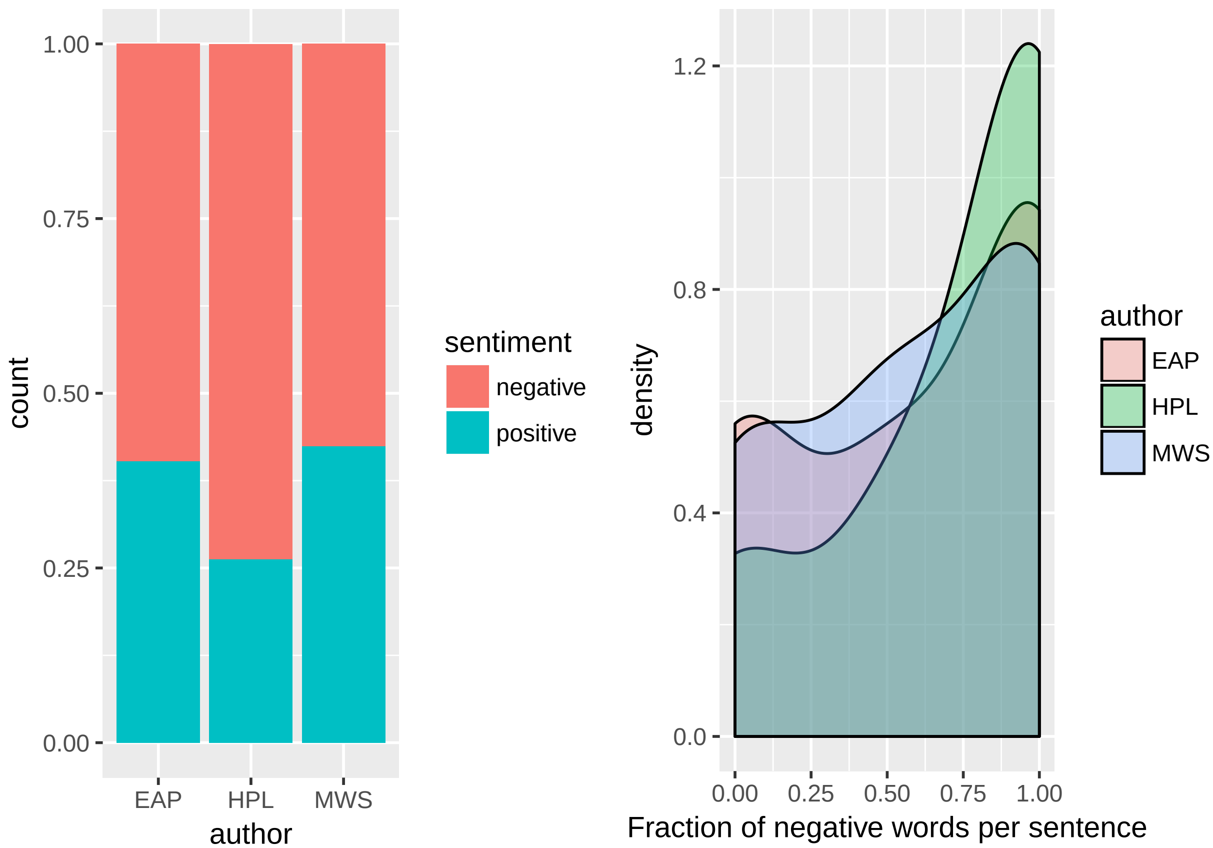

```r
spooky_word %>%
  filter(author == "MWS") %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup() %>%
  group_by(sentiment) %>%
  top_n(10, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to negative/positive sentiment", x = NULL) +
  coord_flip() +
  ggtitle("Sentiment Analysis: MWS")
```



Comparing MWS with EAP and HLP, we can clearly see that her use of positive words is very characteristic of her style. Words such as `love`, `happy`, and `joy` are as frequent as negative words such as `death`, `fear`, or `misery`, which is quite surprising considering the genre of her writing. Also, as we have seen from the analysis of vocabulary, EAP and HPL exhibit distinctive sentiments in their writings as well. Although both of their writings are predominantly negative, EAP's negative sentiment is more obscure and indirect, compared to HPL's (`doubt`, `difficulty`, `lost` vs. `strange`, `terrible`, `hideous`). EAP's positive words also show more refined sentiments with words such as `regard`, `beauty`, or `respect`, which is consistent from other aspects of his writings.

```
p3 <- spooky_word %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  ggplot(aes(author, fill = sentiment)) +
  geom_bar(position = "fill")

p4 <- spooky_word %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  group_by(author, id, sentiment) %>%
  count() %>%
  spread(sentiment, n, fill = 0) %>%
  group_by(author, id) %>%
  summarise(neg = sum(negative),
            pos = sum(positive)) %>%
  arrange(id) %>%
  mutate(frac_neg = neg/(neg + pos)) %>%
  ggplot(aes(frac_neg, fill = author)) +
  geom_density(bw = .2, alpha = 0.3) +
  theme(legend.position = "right") +
```

```r
  labs(x = "Fraction of negative words per sentence")

layout <- matrix(c(1,2),1,2,byrow=TRUE)
multiplot(p3, p4, layout=layout)
```



Overall, HPL's sentiment is the most negative among three, with relatively little use of positive words in his writing. MWS shows the least negative sentiment in general, but EAP also used positive sentiments to maintain more neutral tone of the writing as well.

Conclusion

1. Stylistically, HPL generally uses the longest sentences among three, both in terms of the number of characters and the number of words, and EAP the shortest among three.
2. In the use of vocabulary, EAP uses longer words, compared to the other two, although the difference is minor. MWS dominantly uses female indicators and pronouns in her writing, compared to HPL who rarely uses them in his. The difference between HPL and EAP is also significant.
3. MWS has the most characteristic use of vocabulary of all three, using lots of names and positive words like `life`, `love`, and `heart`. HPL and AEP's frequently used words (`folk`, `bearded`, `brown`, or `attic` vs. `monsieur` or `madame`) also show the clear difference between two writer's preferred settings in their stories. This result is consistent in the analysis of bigrams.
4. MWS's use of lots of positive words create a positive sentiment in her writing, clearly distinguishing her from the other two. HPL's sentiment is the most negative among three, with relatively little use of positive words in his writing. EAP's sentiment is more neutral, and we can further see some differences in how each author's sentiment is manifested.