Spooky Data - Which author wrote this sentence?

Chunzi Wang

Febuary 5, 2018

1 Install needed packages and load libraries

```
packages.used <- c("ggplot2","dplyr","tidytext","wordcloud","stringr","ggridges","tidyr","reshape2","to

# 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(tidytext)
library(wordcloud)
library(stringr)
library(ggridges)
library(tidyr)
library(reshape2)
library(topicmodels)
library(treemapify)
```

2 Import data

```
spooky <- read.csv('E:\\GitHub\\spring2018-project1-chunziwang\\data\\spooky.csv', as.is = TRUE)

dim(spooky)
```

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

```
summary(spooky)
```

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

Spooky data has 19579 rows * 3 columns. Each row identifies a sentence text, the author who wrote the sentence, and the unique id associated with the sentence. There're 3 authors in total: Edgar Allen Poe, Mary Shelley, and HP Lovecraft. We can find out how many sentences each author has in this dataset:

```
spooky %>%
  count(author)
```

```
## # A tibble: 3 x 2
##   author     n
```

```
##    <chr> <int>
## 1    EAP  7900
## 2    HPL  5635
## 3    MWS  6044
```

Edgar Allen Poe has 7900 sentences, Mary Shelley has 6044 sentences and HP Lovecraft has the least, 5635 sentences.

To prepare the data for next-step analysis, I changed the author column into factor, unnested spooky text into tokens and removed the stop words since I believe it'd be much more interesting to look at the unique words horror fiction writers commonly use.

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

spooky_words <- spooky %>%
  unnest_tokens(word,text)

spooky_words <- spooky_words %>%
  anti_join(stop_words)
```

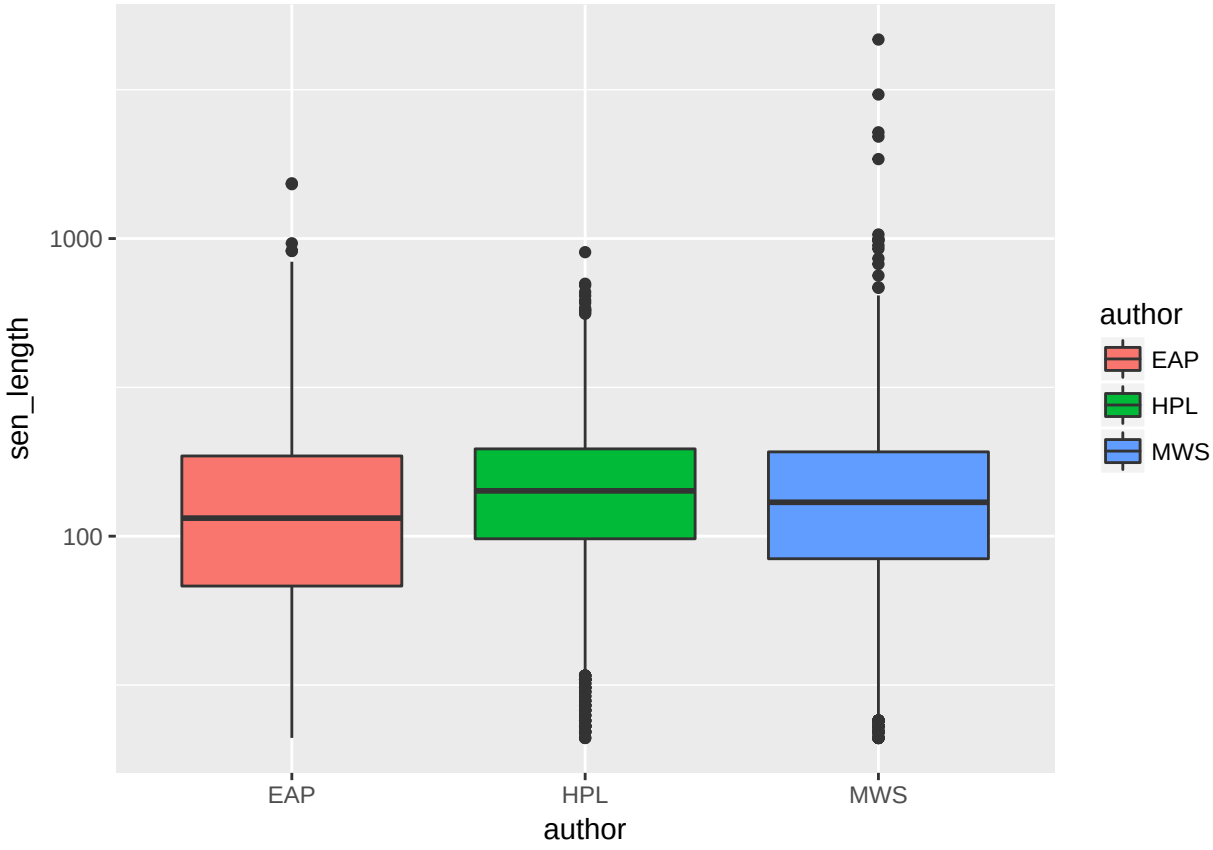
3 Exploratory data analysis

The main goal here is to find as many differences among 3 authors in their word choice, rhetoric, writing style, literary production, and so on to help extract features that could help differentiate one author from another just by looking at their texts.

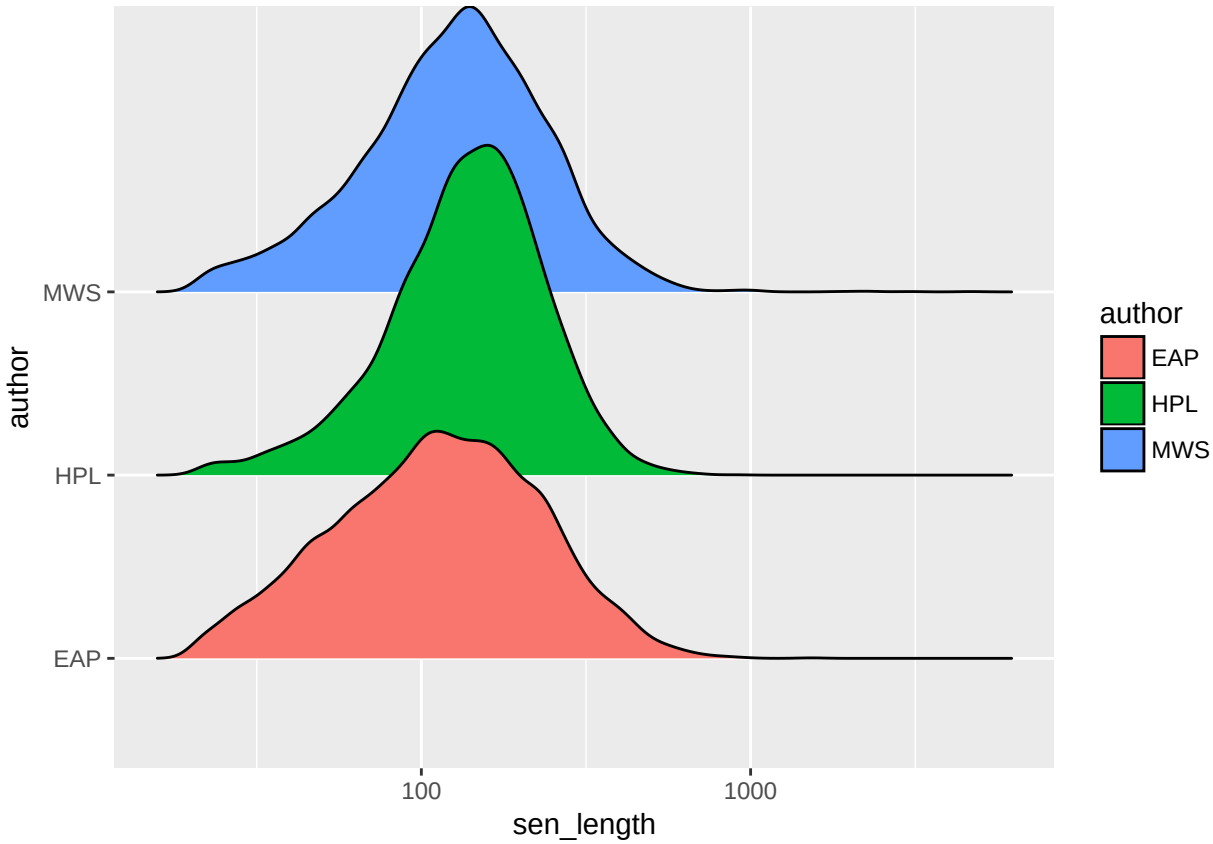
Sentence length distribution of each author

At first, I want to look at if there's any author who often writes longer or shorter sentences compared to others.

```
spooky$sen_length <- str_length(spooky$text)

# boxplot
spooky %>%
  ggplot(aes(x=author,y=sen_length,fill=author)) +
  geom_boxplot() +
  scale_y_log10()
```




```
# density ridge plot
spooky %>%
  ggplot(aes(x=sen_length,y=author,fill=author)) +
  geom_density_ridges() +
  scale_x_log10()
```

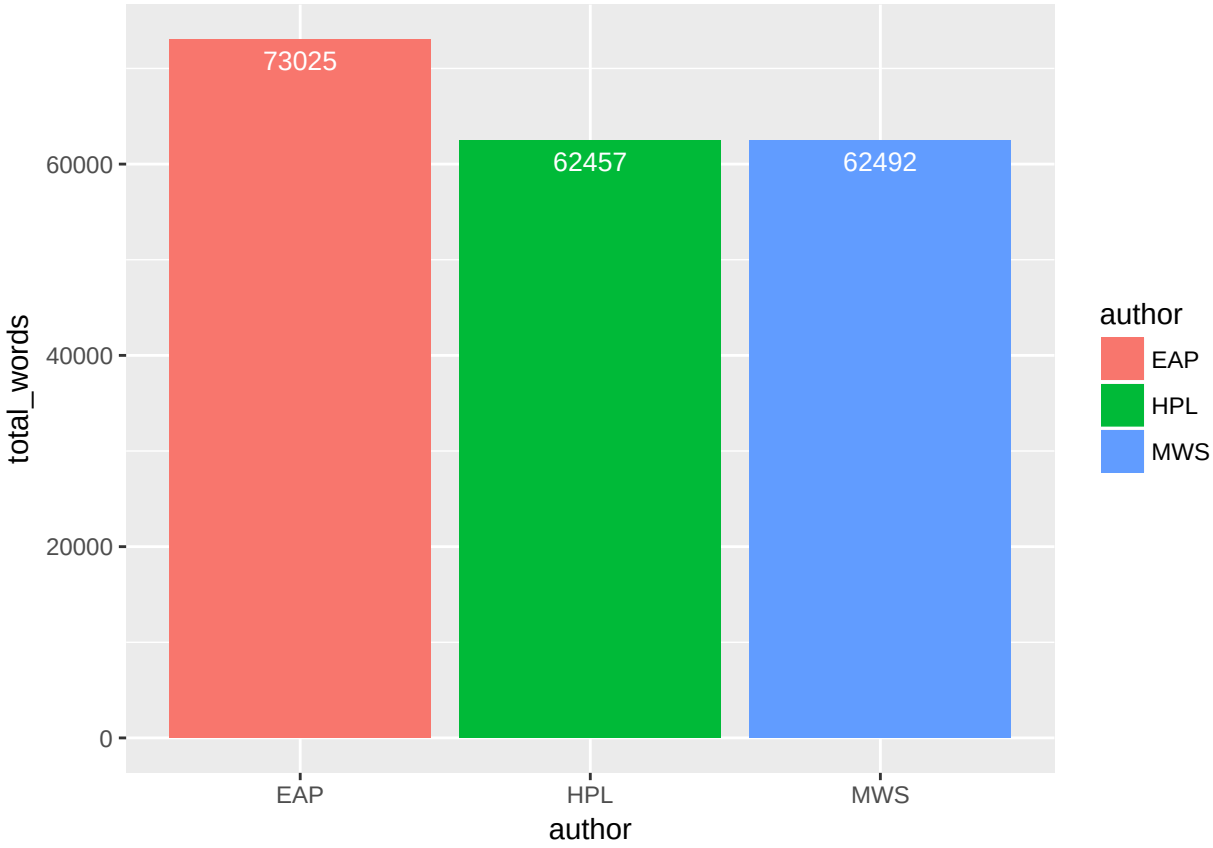


From the boxplot, it's clear that Allen Poe's sentences tend to be shorter since almost half of them fall below 100 characters per sentence while HP Lovecraft has 75% of his sentences longer than 100 characters per sentence. Mary Shelly's sentence length comes between Allen Poe and HP Lovecraft but she owns most of the extreme values - sentences that's over 1000 characters!

From the density plot, the peak of HP Lovecraft and Mary Shelly is slightly right-skewed compared with Allen Poe, indicating higher percentage of their sentences is longer. The slope of ridge of HP Lovecraft below 100 is comparatively flat and makes his whole distribution narrower than the other two.

Count how many meaningful words each author used

```
spooky_words %>%
  count(author) %>%
  rename (total_words=n) %>%
  ggplot(aes(x=author,y=total_words,fill=author)) +
  geom_col() +
  geom_text(aes(label=total_words), vjust=1.6, color="white", position=position_dodge(0.9),size=3.5)
```

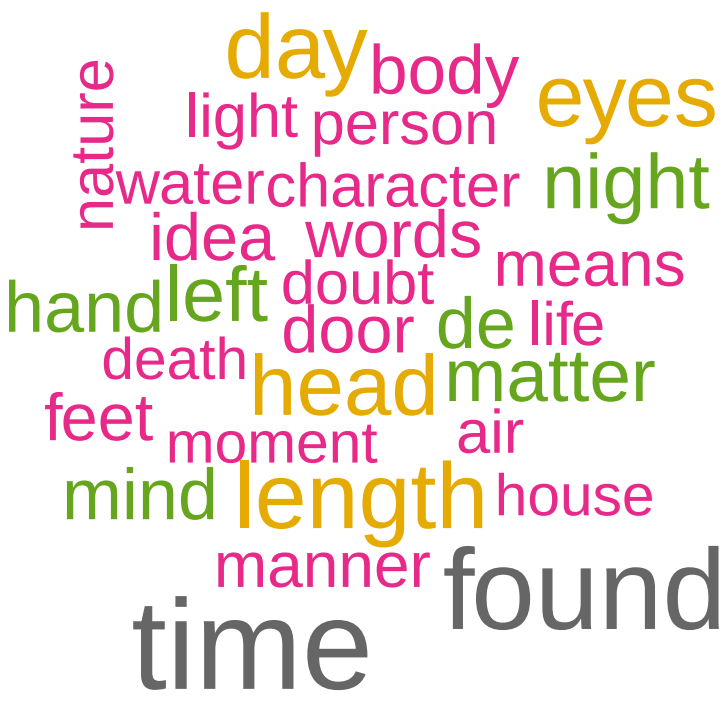


Stop words aside, Allen Poe has used around 73000 words in toal. Both HP Lovecraft and Mary Shelly used 62000 words. Since there're 1000 more sentences of Allen Poe's in the spooky dataset it makes sense that Allen Poe used more words.

What're the most frequent words each author used

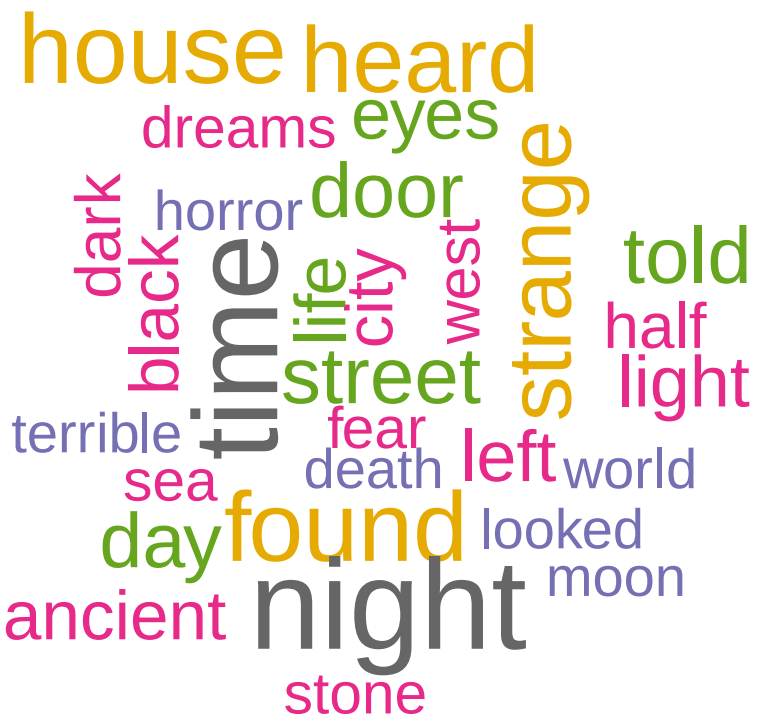
To get a first impression of the most commonly used words by each author, I made a word cloud of the top 30 words here for each author. This is Allen Poe. I found he likes detailed depiction of person since words like "eyes", "head", "hand", "mind", "body", "manner", "character", "person", "words" are in high frequency here.

```
spooky_words %>%
  filter(author=="EAP") %>%
  count(word) %>%
  arrange(desc(n)) %>%
  with(wordcloud(word,n,max.words=30,colors=brewer.pal(8, "Dark2")))
```



In comparison, HP Lovecraft prefers depicting things/scenes more. He used adjectives like “strange”, “ancient”, “dark”, “terrible”, and words such as “house”, “street”, “stone”, “sea”, and “moon”. These words seem to show the background scene of his stories. Typical spooky words such as “fear”, “horror”, “death” also show up here, indicating HP Lovecraft’s frequent uses of them and it makes perfect sense for a horror fiction writer.

```
spooky_words %>%
  filter(author=="HPL") %>%
  count(word,sort=TRUE) %>%
  with(wordcloud(word,n,max.words=30,colors=brewer.pal(8, "Dark2")))
```



Mary Shelly's different in that her words evolve more emotion and are more passionate. Probably it's the different characteristics of female writers. Character's names such as "raymond", "idris", "adrian", "father" and "friend" take place quite often. She depicted "love", "heart", "soul", "spirit" but also heavy "death". There're "tears", "fear", and "hope". From these words I could tell she's a very interesting writer becuase of her heavy usage of Jane-Austen-style romance words while also remaining the essence of a horror fiction.

```
spooky_words %>%
  filter(author=="MWS") %>%
  count(word,sort=TRUE) %>%
  with(wordcloud(word,n,max.words=30,colors=brewer.pal(8, "Dark2")))
```

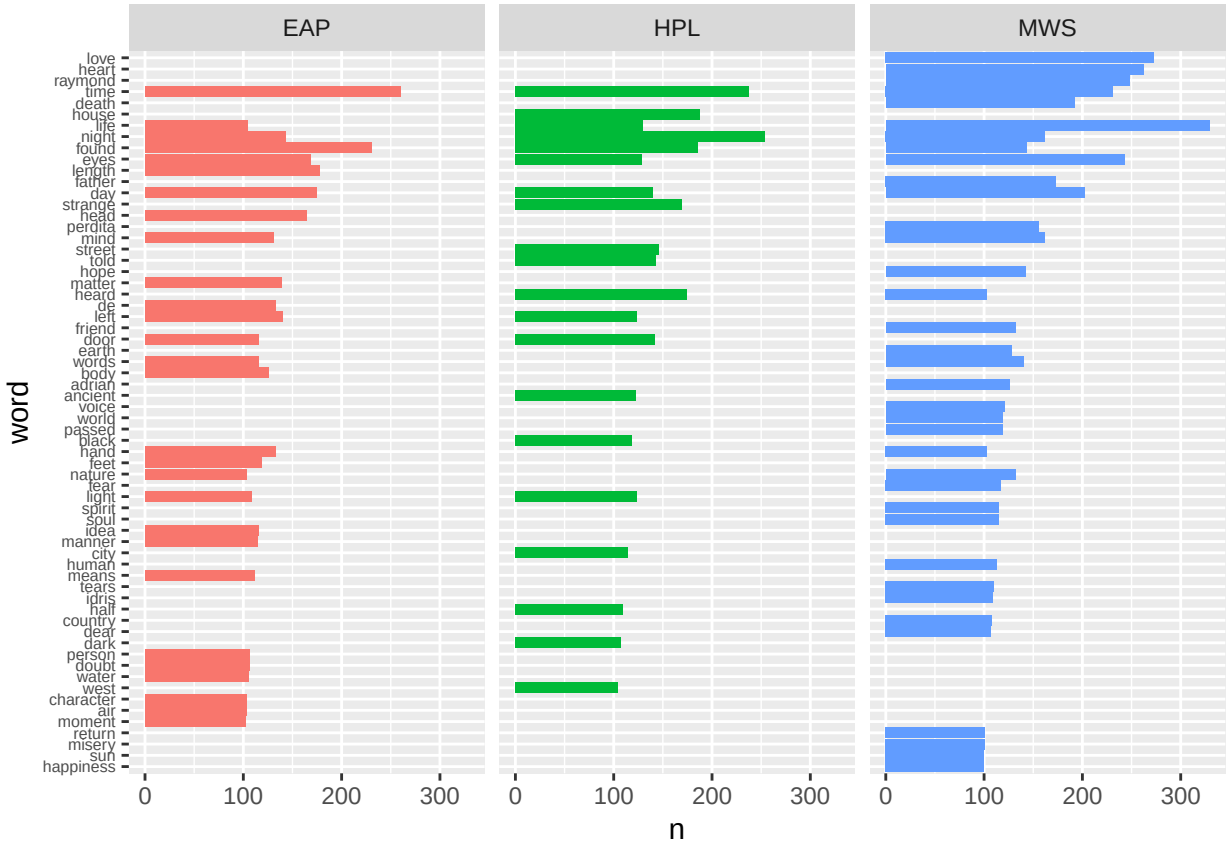
I found nouns like “life”, “night”, “time” and verbs like “found”, “heard” are commonly used by all 3 authors. It’s understandable since they are writing in the same genre where “life” and “night” are often talked about, “found” and “heard” often suggest something unusal happened.

To be more clear, I made a bar plot to show the words and their frequency by author. I only used the first 84 rows in the dataframe because only these rows have a seperate word count >= 100.

```
word_counts_by_author <- spooky_words %>%
  count(author,word,sort=TRUE)

head(word_counts_by_author)
```

```
## # A tibble: 6 x 3
##   author    word     n
##   <fctr>   <chr> <int>
## 1    MWS    life   329
## 2    MWS    love   273
## 3    MWS   heart   262
## 4    EAP    time   260
## 5    HPL   night   254
## 6    MWS raymond   248
```

```
ggplot(word_counts_by_author[1:84,],aes(x=reorder(word,n),y=n,fill=author)) +
  geom_col(show.legend=FALSE) +
  coord_flip() +
  facet_wrap(~author) +
  xlab("word") +
  theme(axis.text.y=element_text(size=6))
```



This comparison bar plot did a good job in showing the difference of high-frequency words every author used. Their personal preference in word choice is completely different. There're some words like "love", "heart", "hope", "world", "sun", "happiness" that Mary Shelly used a lot but hardly appear in other author's works, and vice versa.

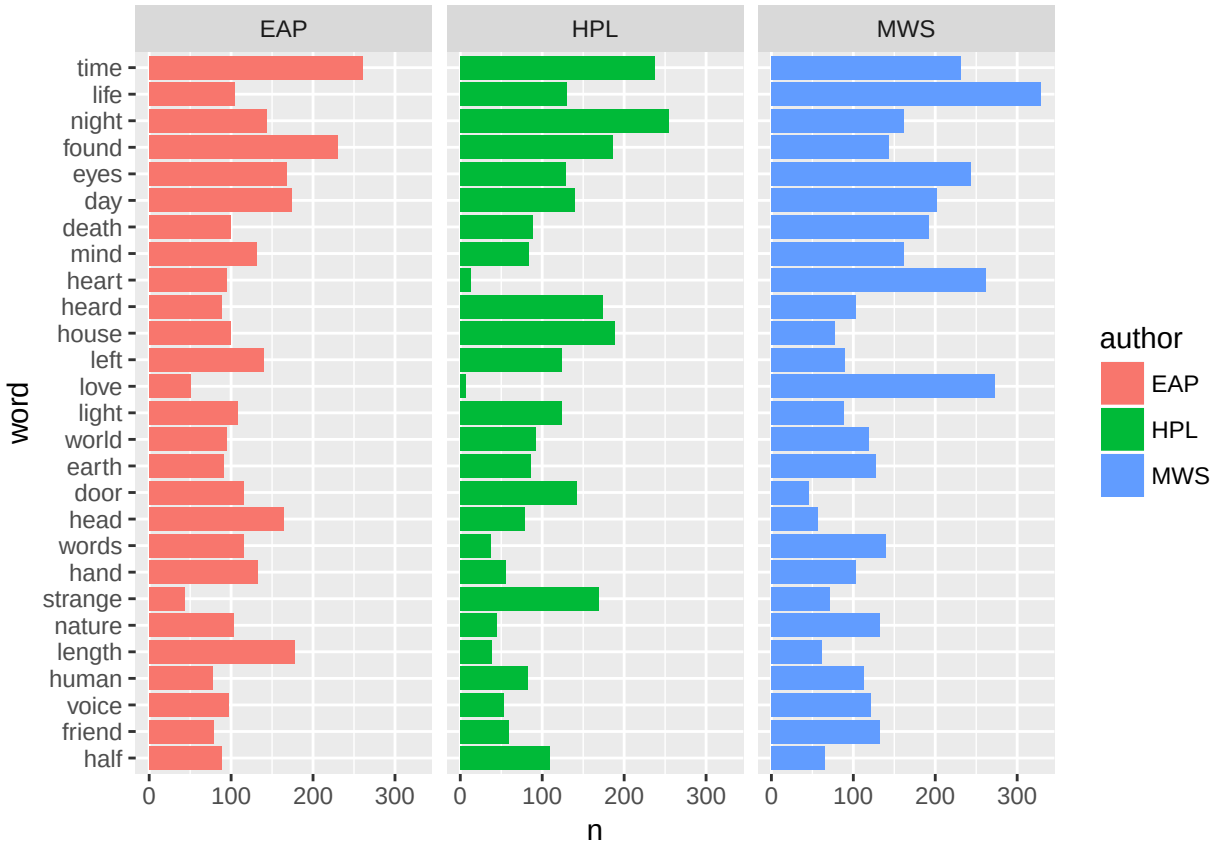
The problem is that for some words such as "house" and "death" also appear comparatively frequently in all 3 authors'work, but due to the seperate word count by author is below 100, it doesn't show up in this bar plot, and that leaves a misleading message that shows Allen Poe and HP Lovecraft didn't use the word "death", which it's not true. So we need to make another bar plot to contain the frequent words that are commonly used in all three authors' works.

```
word_counts <- spooky_words %>%
  count(word,sort=TRUE)

head(word_counts)
```

```
## # A tibble: 6 x 2
##    word     n
##   <chr> <int>
## 1  time   729
## 2  life   563
## 3 found   559
## 4 night   559
## 5  eyes   540
## 6   day   516
```

```
word_counts %>%
  left_join(word_counts_by_author,by="word") %>%
```

```
  head(81) %>%
  ggplot(aes(x=reorder(word,n.y),y=n.y,fill=author)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~author) +
  xlab("word") +
  ylab("n")
```

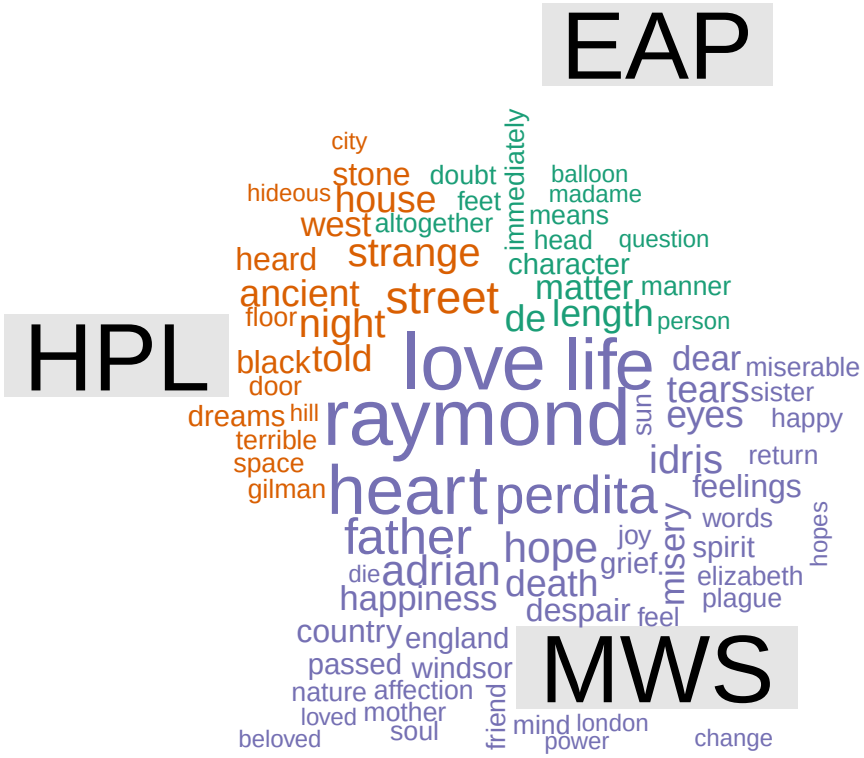


These are the overall most popular words by all three authors. We can see that for high-frequency words on the top such as "time", "life", "night", "found", "eyes", "day", and "death" are almost equally frequent for all authors. Words that shows affection in meaning such as "heart" and "love" are more frequently used by Mary Shelly, then Allen Poe but seldom occur in HP Lovecraft's writing. Words that are more frequently used by HP Lovecraft are "strange" and "night".

Comparison cloud

Comparison cloud hides the common words that every author has in common and only shows the unique words they used, at least in the top 80 words ranked by their count.

```
spooky_words %>%
  count(author,word,sort=TRUE) %>%
  acast(word~author,value.var="n",fill=0) %>%
  comparison.cloud(colors=brewer.pal(8, "Dark2"),max.words=80,scale=c(2.5,0.2))
```



For the top 80 words, more than half of them belongs to Mary Shelly. It's clear to see three authors' differences in that Mary Shelly used lots of words that is connected to feelings and emotions, positive and negative, such as "love", "heart", "tears", "misery", "happiness", "affection", "despair" and so on. HP Lovecraft wrote more about things and environment - "street", "night", "strange", "house" etc. There're fewer Allen Poe's words here and they doesn't look very outstanding. Some are connected to human - "person", "manner", "character", "head", "madame", and others feel a little strange in showing up here such as "balloon", "immediately" and "altogether".

From this process, we could get a general sense of each author's writing style and their preferred choice of words. Although all writing horror fictions, they tend to tell the stories differently. We're going to spot more difference in the following sentiment analysis.

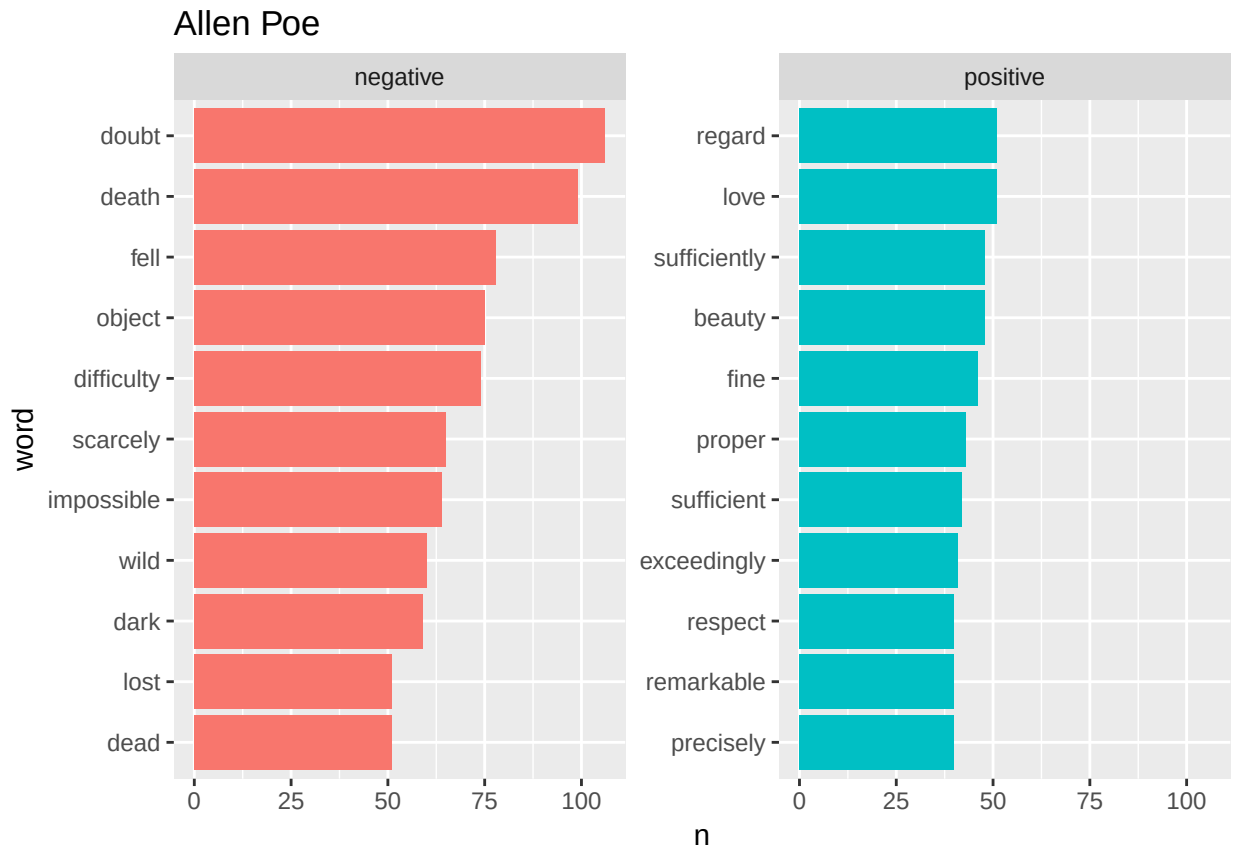
4 Sentiment analysis

The basic sentiments in sentiment analysis is "positive" and "negative". Since spooky data is about texts extracted from horror fiction, the assumption is that there'll be a noticeably higher proportion of words that are identified as negative than positive. This is just my personal guess. Let's look at what negative and positive words each author used based on "bing" lexicon.

Top10 most frequent positive and negative words used by each author

```
spooky_words %>%
  inner_join(get_sentiments("bing")) %>%
  count(author,word,sentiment,sort=TRUE) %>%
  filter(author=="EAP") %>%
```

```
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(x=reorder(word,n),y=n,fill=sentiment)) +
  geom_col(show.legend=FALSE) +
  coord_flip() +
  facet_wrap(~sentiment,scales="free_y") +
  xlab("word") +
  ggtitle("Allen Poe")
```

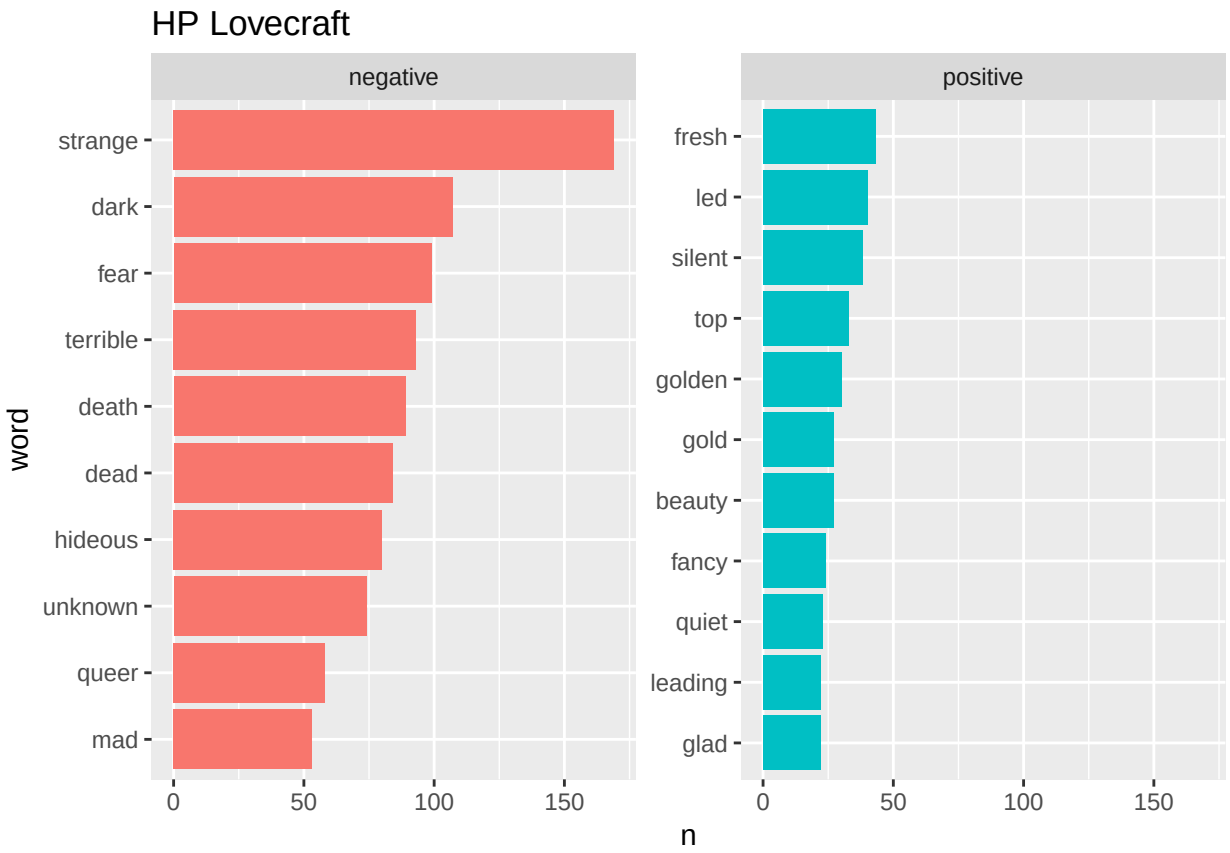


I found:

- Allen Poe used much more negative words than positive words, but he used "love" too.
- Compared with Mary Shelly, his negative words such as "doubt", "object", "scarcely" and positive words such as "regard", "sufficiently", "exceedingly", and "precisely" are much more unpredictable and interesting.

```
spooky_words %>%
  inner_join(get_sentiments("bing")) %>%
  count(author,word,sentiment,sort=TRUE) %>%
  filter(author=="HPL") %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(x=reorder(word,n),y=n,fill=sentiment)) +
  geom_col(show.legend=FALSE) +
```

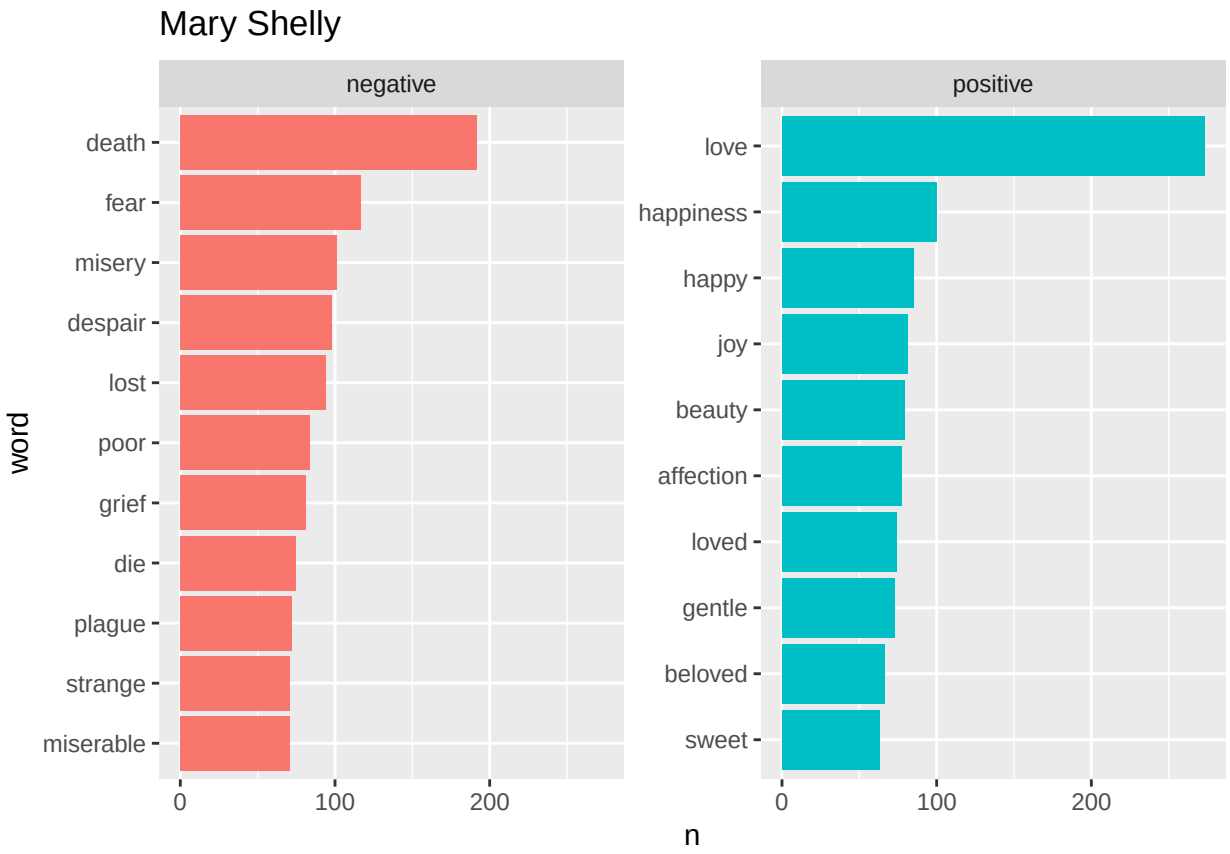
```
  coord_flip() +
  facet_wrap(~sentiment,scales="free_y") +
  xlab("word") +
  ggtitle("HP Lovecraft")
```



I found:

- I'd say HP Lovecraft is most stingy about using positive words.
- The positive words seem a little odd: "fresh", "silent", "golden", "quiet", "leading"... It certifies HP Lovecraft doesn't really use commonly recognized positive words so that words that sound neutral are being clustered into the positive group.

```
spooky_words %>%
  inner_join(get_sentiments("bing")) %>%
  count(author,word,sentiment,sort=TRUE) %>%
  filter(author=="MWS") %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(x=reorder(word,n),y=n,fill=sentiment)) +
  geom_col(show.legend=FALSE) +
  coord_flip() +
  facet_wrap(~sentiment,scales="free_y") +
  xlab("word") +
  ggtitle("Mary Shelly")
```



I found:

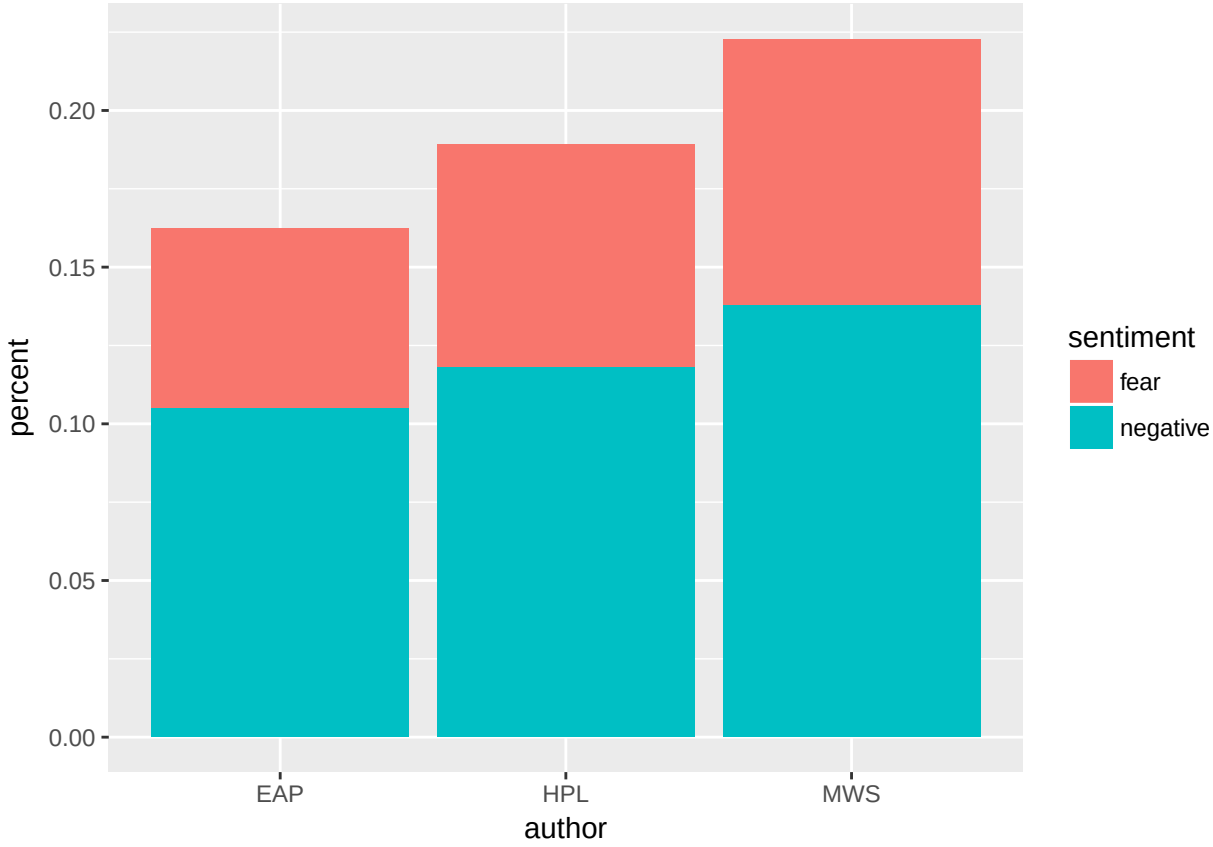
- Mary Shelly used the same amount of positive words and negative words.
- Unlike the other two authors, her positive and negative words are very pure and predictable. No matter "love", "happiness", "joy", "beauty", "sweet", or "death", "fear", "misery", "despair", "grief", "plague", these are all strong positive or negative words that are most commonly associated with positive or negative feelings and emotions.

Which author used the most negative/fear words

In another lexicon "nrc", sentiments have been classified into 10 catogories: trust, surprise, fear, negative, sadness, anger, positive, joy, disgust, anticipation. Since spooky is associated with this dreadful, horrible feeling, I want to look at which author used the most negative words and words associated with fear.

```
spooky_sentiment <- spooky_words %>%
  group_by(author) %>%
  mutate(author_total=n()) %>%
  ungroup() %>%
  inner_join(get_sentiments("nrc"))

spooky_sentiment %>%
  count(author,sentiment,author_total) %>%
  mutate(percent=n/author_total) %>%
  filter(sentiment %in% c("negative","fear")) %>%
  arrange(percent) %>%
```

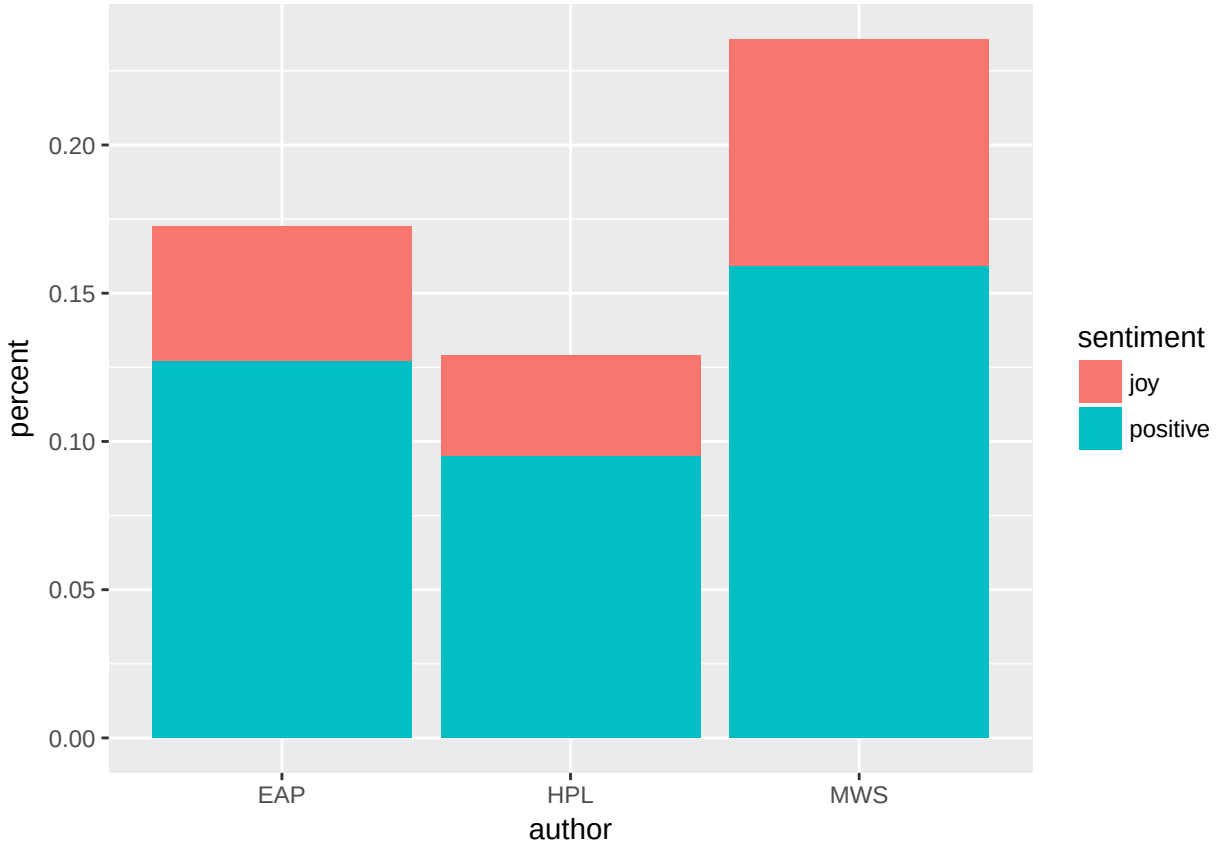
```
  ggplot(aes(x=author,y=percent,fill=sentiment)) +
  geom_col()
```



To my surprise, Mary Shelly used the most negative words (14%) and words associated with fear (8%), so combined it's 22%, almost a quarter of the whole texts. Before the plotting I thought HP Lovecraft would come on top since from the analysis above he seems to tend to use more negative words. But after second thought, it makes sense in that the negative words in "bing" lexicon for HP Lovecraft could be classified into other non-negative or non-fear categories in "nrc" lexicon.

Let's look at positive sentiment and sentiment associated with joy with the same method.

```
spooky_sentiment %>%
  count(author,sentiment,author_total) %>%
  mutate(percent=n/author_total) %>%
  filter(sentiment %in% c("positive","joy")) %>%
  arrange(percent) %>%
  ggplot(aes(x=author,y=percent,fill=sentiment)) +
  geom_col()
```



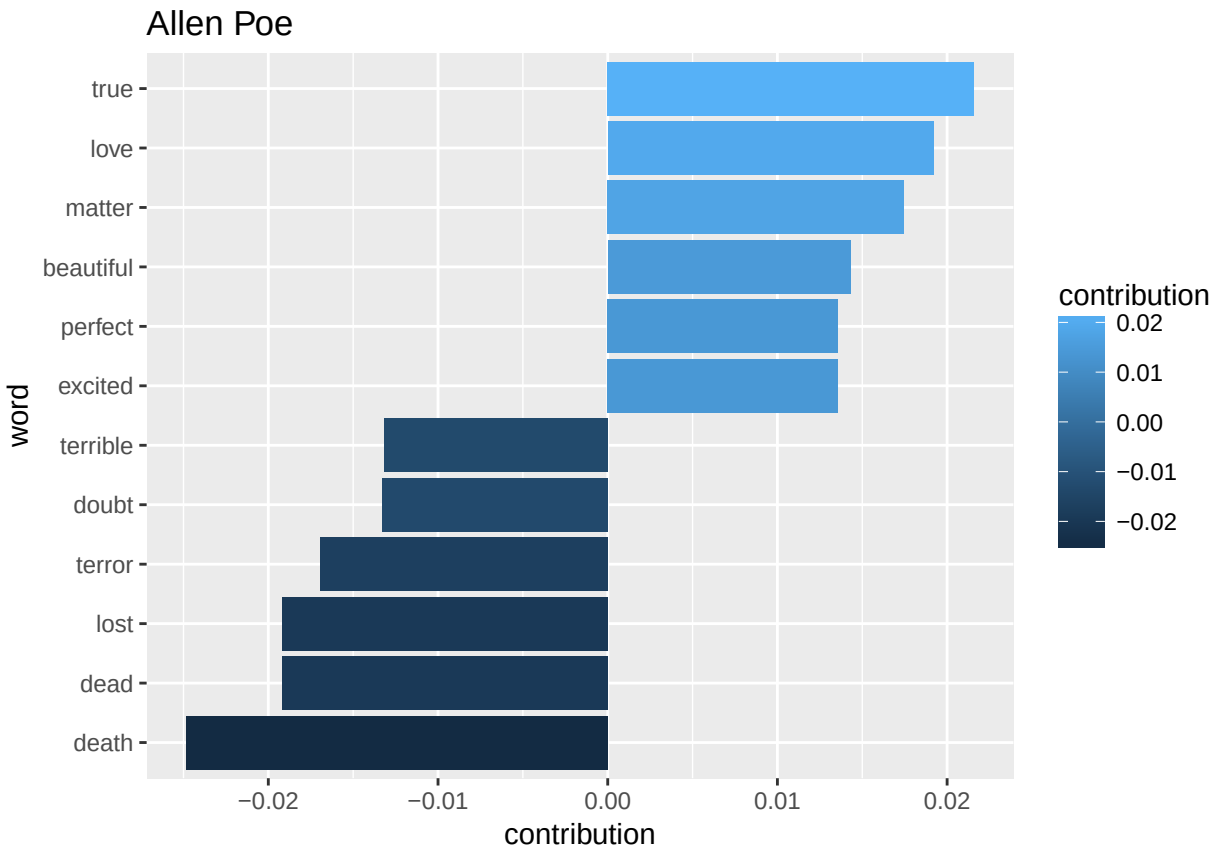
This time, Mary Shelly still comes on top: 16% of her words are positive and 7.5% are associated with joy. HP Lovecraft is the lowest, only 12.5% of his words are positive or associated with joy, which is consistent with what we have found during the journey.

Give sentiment a score

In another lexicon "afinn", each sentiment comes with a score range from -5 to 5. A negative number means negative sentiment and the number shows how significant it is to the sentiment. For example abandon scores -2 while abhor scores -3. I calculated the score multiply the times a word appear in the data and then divided by the total number of words for each author to see the sentiment contributions by individual words.

```
sentiment_contributions <- spooky_words %>%
  count(author,word,sort=TRUE) %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(author) %>%
  mutate(contribution = score*n/sum(n)) %>%
  ungroup()

sentiment_contributions %>%
  filter(author=="EAP") %>%
  arrange(desc(abs(contribution))) %>%
  head(12) %>%
  ggplot(aes(x=reorder(word,contribution),y=contribution,fill=contribution)) +
  geom_col() +
  coord_flip() +
  xlab("word") +
```

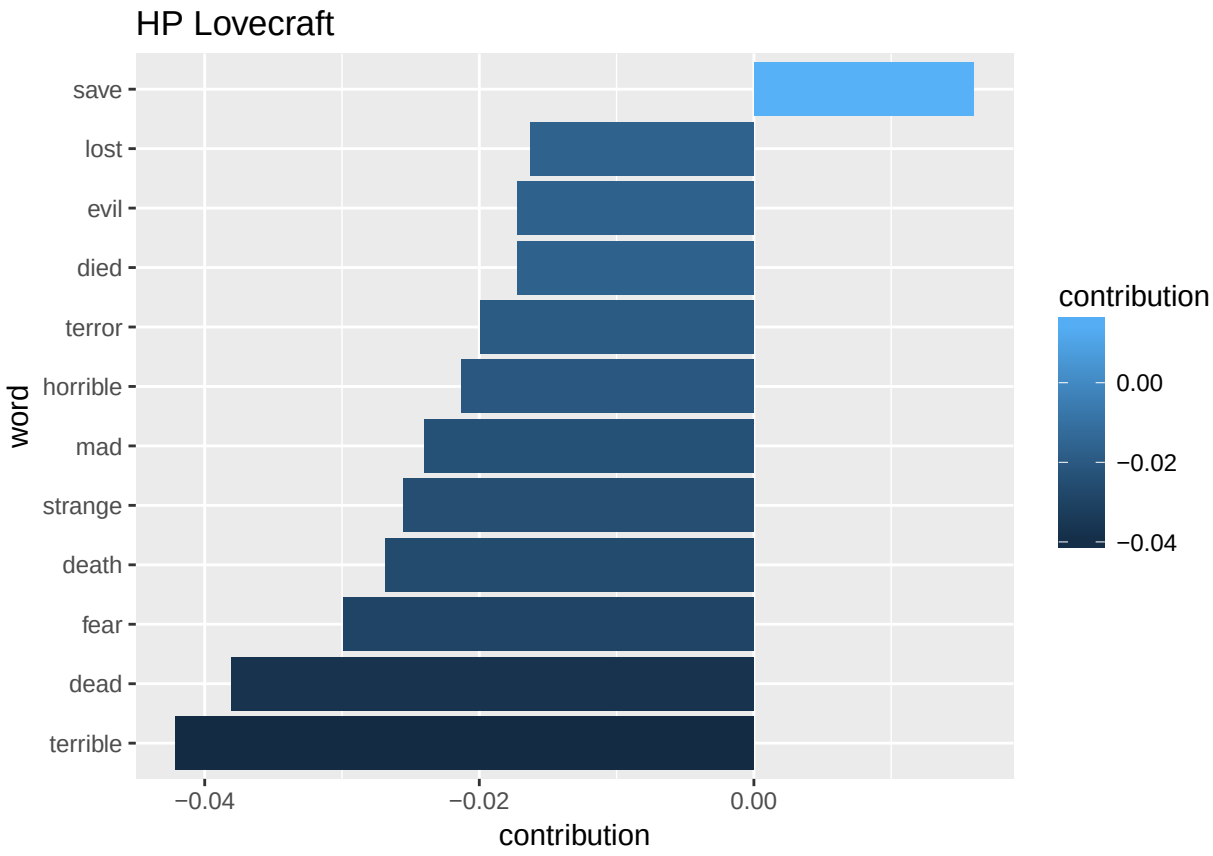
```
  ggtitle("Allen Poe")
```



I found:

- Both positive and negative words make equal sentiment contribution.
- It's interesting that "True" is the positive word that makes the highest contribution in Allen Poe's texts.

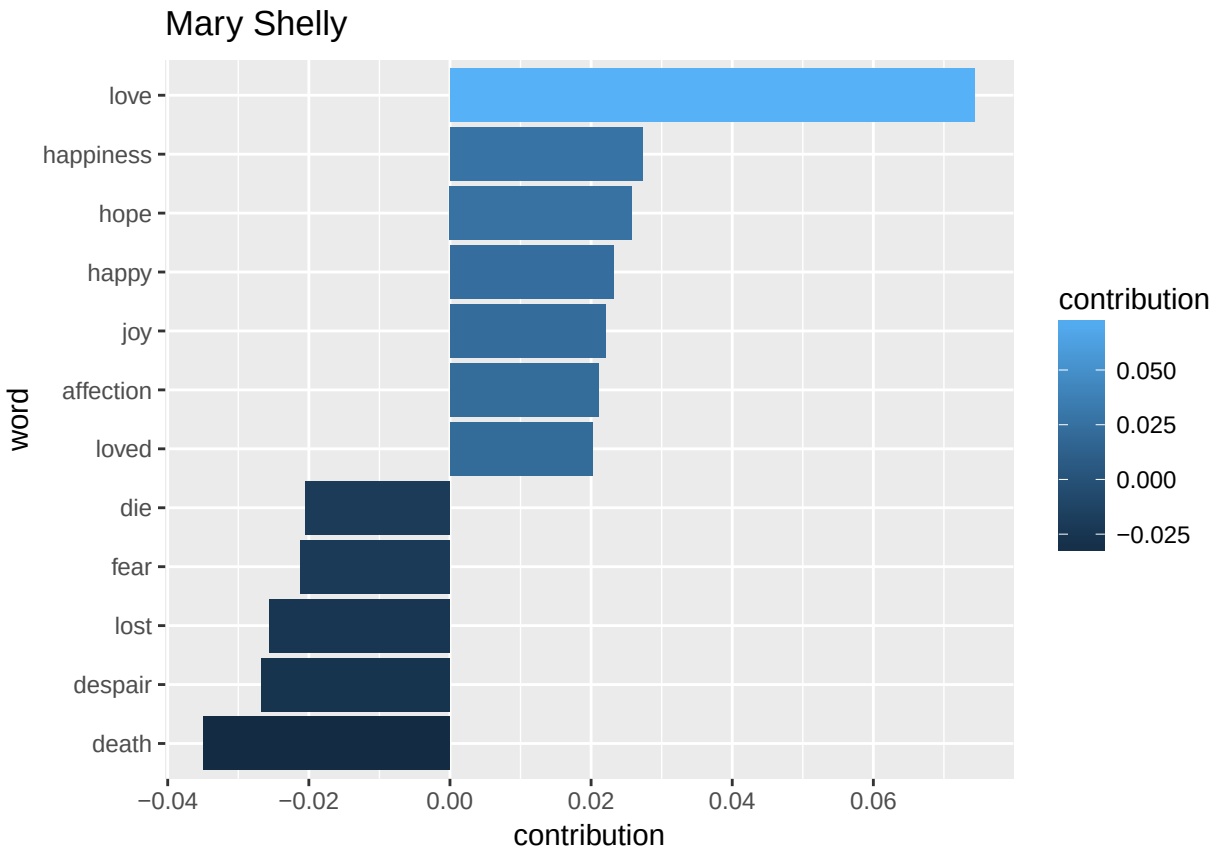
```
sentiment_contributions %>%
  filter(author=="HPL") %>%
  arrange(desc(abs(contribution))) %>%
  head(12) %>%
  ggplot(aes(x=reorder(word,contribution),y=contribution,fill=contribution)) +
  geom_col() +
  coord_flip() +
  xlab("word") +
  ggtitle("HP Lovecraft")
```



I found:

- 11 out of 12 words that make a high contribution to HP Lovecraft's texts are negative. Negative words do count more in HP Lovecraft's works.
- After taking the sentiment score into consideration, "terrible" becomes the negative word that has the highest contribution . "strange" descends to No.5 instead of the No.1 in the most frequently used negativ words from previous plots.

```
sentiment_contributions %>%
  filter(author=="MWS") %>%
  arrange(desc(abs(contribution))) %>%
  head(12) %>%
  ggplot(aes(x=reorder(word,contribution),y=contribution,fill=contribution)) +
  geom_col() +
  coord_flip() +
  xlab("word") +
  ggtitle("Mary Shelly")
```



I found:

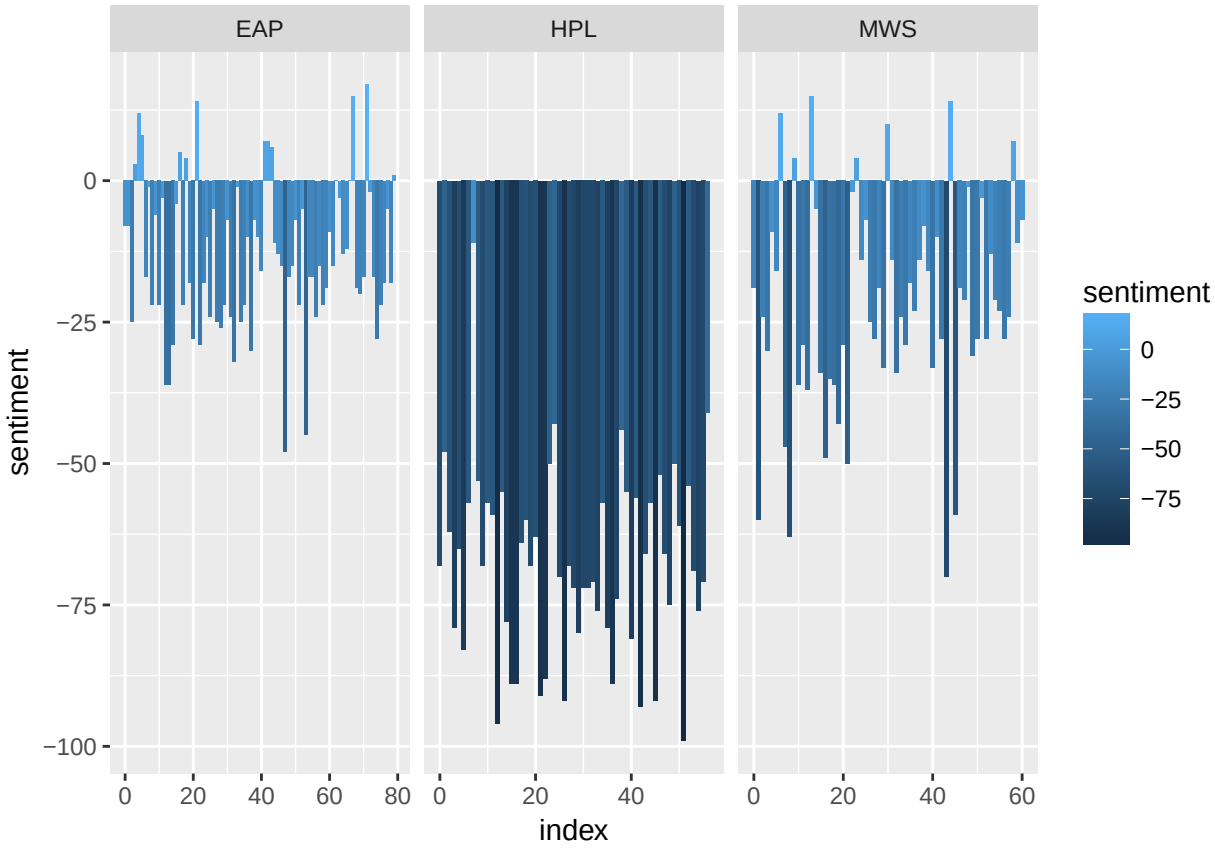
- Positive words have a higher contribution than negative words for Mary Shelly. Especially for "love".
- Lots of words that have a high contribution could be stemmed: "love" and "loved", "happiness" and "happy", "death" and "die". It shows Mary often uses this line of words.
- For all of the three authors, "death", "dead", and "die" are always the top negative words that has the highest contribution. I'd think it's in response to the spooky theme of their texts.

Sentiment in chunks

So far I only did sentiment analysis based on single words. In a sentence, there're usually a mixture of positive and negative words but a reader could always tell whether the sentence as a whole is presenting a positive or negative sentiment. That's what we're going to do here. Since each author has thousands of sentences in the spooky dataset, to prevent the results from looking noisy, I put 100 sentences in a chunk and to see the overall sentiment each chunk presents.

```
# add a column named linenumber to know which line a specific word came from
spooky_words_line <- spooky %>%
  group_by(author) %>%
  mutate(linenumber=row_number()) %>%
  unnest_tokens(word,text) %>%
  ungroup()

# use index to section together lines of text to look at the sentiment of chunks of text
spooky_words_line %>%
```

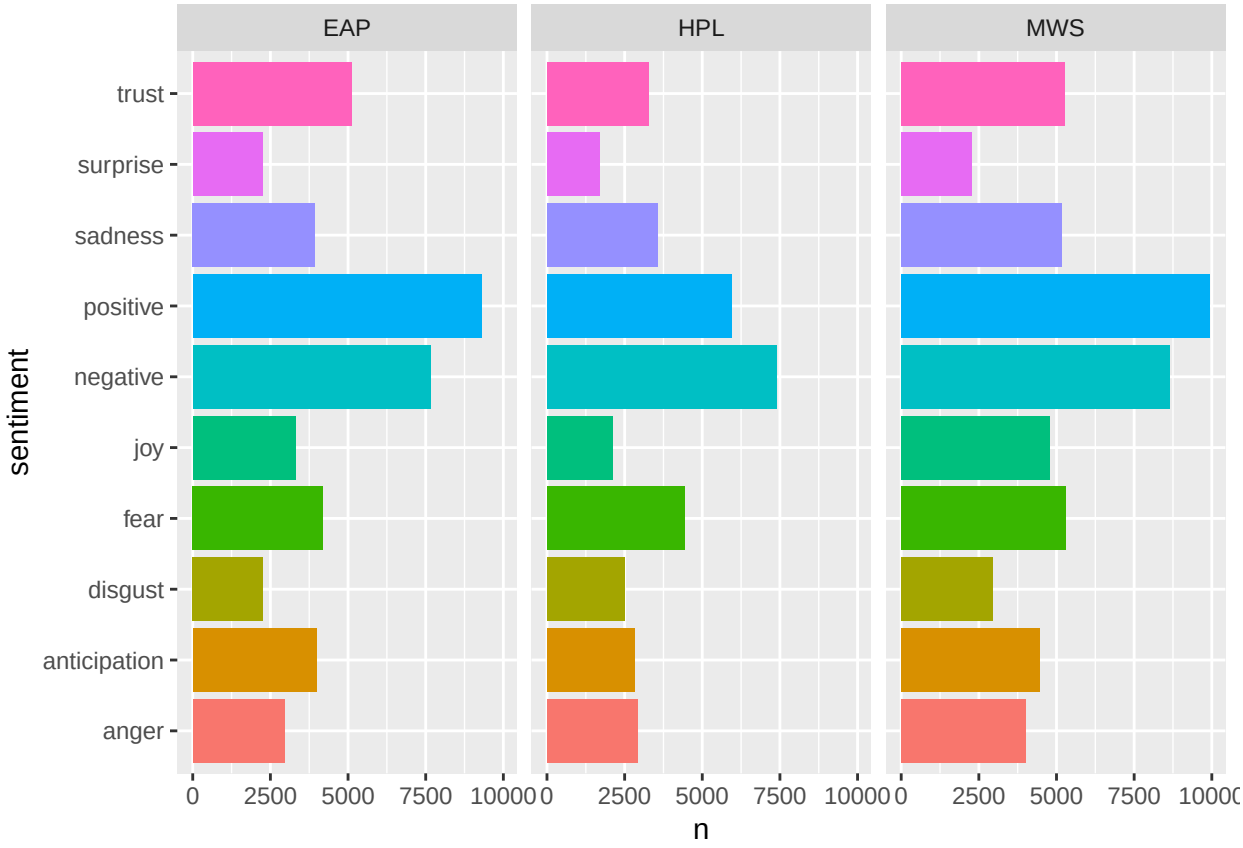
```
  inner_join(get_sentiments("bing")) %>%
  count(author,index=linenumber %/% 100,sentiment) %>%
  spread(sentiment,n,fill=0) %>%
  mutate(sentiment=positive-negative) %>%
  ggplot(aes(x=index,y=sentiment,fill=sentiment)) +
  geom_col() +
  facet_wrap(~author,scales="free_x")
```



Wow! This plot itself looks a little spooky. All of HP Lovecraft's text chunks are deep in negative sentiment. Both Allen Poe and Mary Shelly have majority of chunks in negative but have a few positive chucks every now and then. There're some deep negative chunks in Mary Shelly's work too and that explains her frequent used of negative words we analyzed earlier.

All sentiments for every author

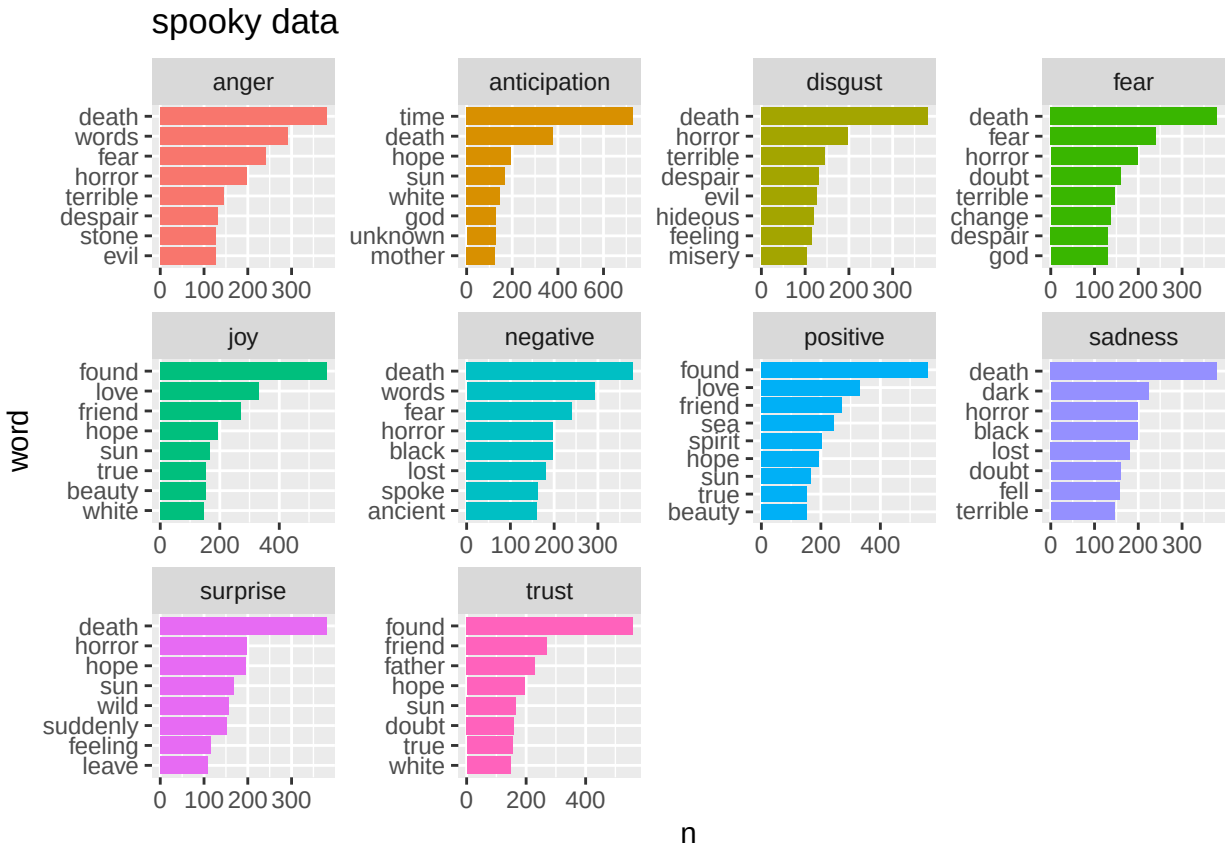
```
spooky_sentiment %>%
  count(author,sentiment) %>%
  ggplot(aes(x=sentiment,y=n,fill=sentiment)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~author) +
  theme(legend.position = "none")
```



This bar plot shows how many words each sentiment contains. Mary Shelly seems to have most words in almost every sentiment category. Her work must be a full of emotional ups and downs. Allen Poe and HP Lovecraft is pretty similar in this sentiment distribution except that Allen Poe has more words associated with trust, positive, surprise and anticipation.

Look at what words comprise each sentiment

```
spooky_sentiment %>%
  count(word,sentiment) %>%
  group_by(sentiment) %>%
  top_n(8) %>%
  ungroup() %>%
  ggplot(aes(x=reorder(word,n),y=n,fill=sentiment)) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~sentiment,scales="free") +
  coord_flip() +
  xlab("word") +
  ggtitle("spooky data")
```



Most of the words under each sentiment segment makes sense, except for "doubt" being placed under trust, and "spoke" under negative. Some words belong to more than one categoty, such as "words", "death", "found" and "god".

5 Topic modeling

Every document is a mixture of topics, and every topic is a mixture of words. I want to see what're the topics that comprise each author's work. I'm going to use LDA (Latent Dirichlet Allocation) here to find the mixture of words that is associated with each topic, and the mixture of topics that describes each document.

Turn dataframe into DocumentTermMatrix

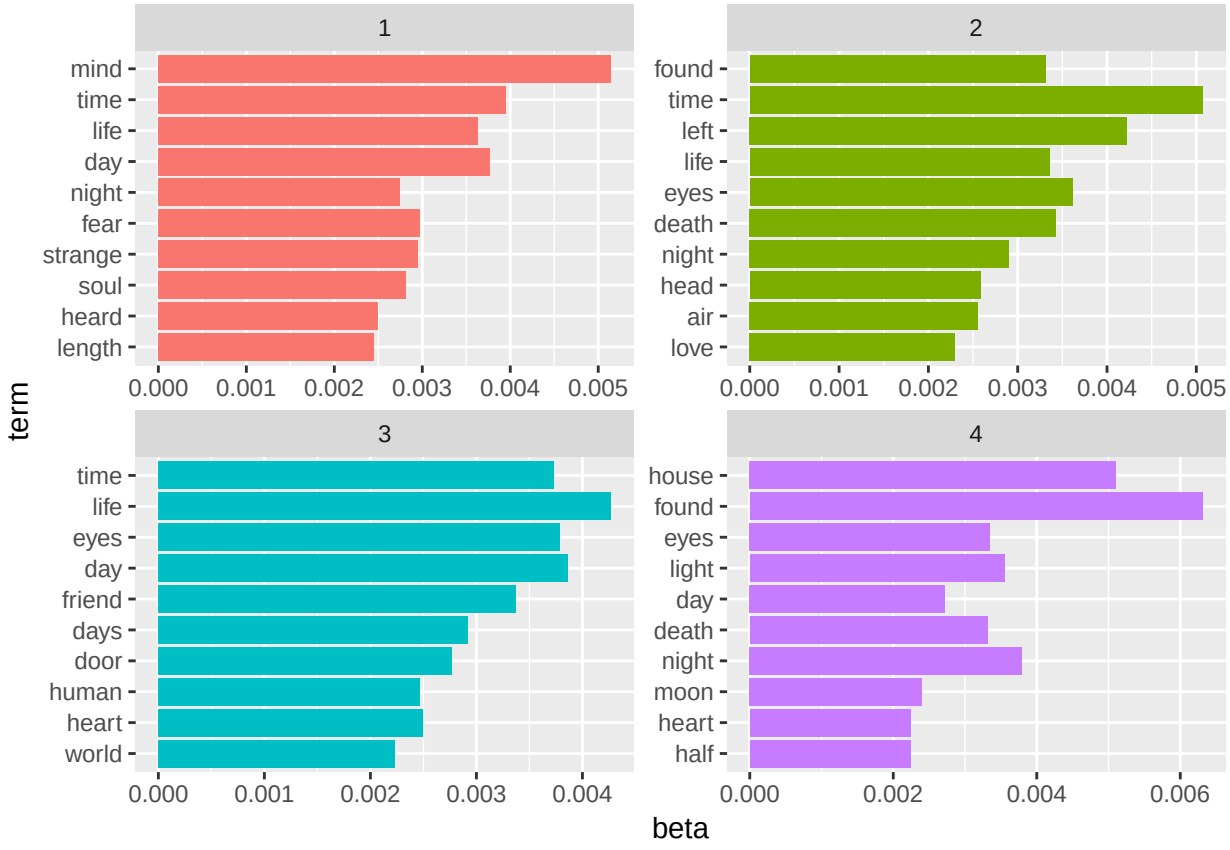
```
freq <- spooky_words %>%
  count(id,word)

# cast the dataframe into a DTM
spooky_words_dtm <- freq %>%
  cast_dtm(id,word,n)
```

Every topic is a mixture of words

Based on the previous analysis I decided to set the k value to 4, which means to create a 4 topic LDA model. I tried 8 topics first but found the words in each topic are pretty similar. So I changed it into 4 in order to

differentiate each topic. Here're the words that are grouped into each topic.

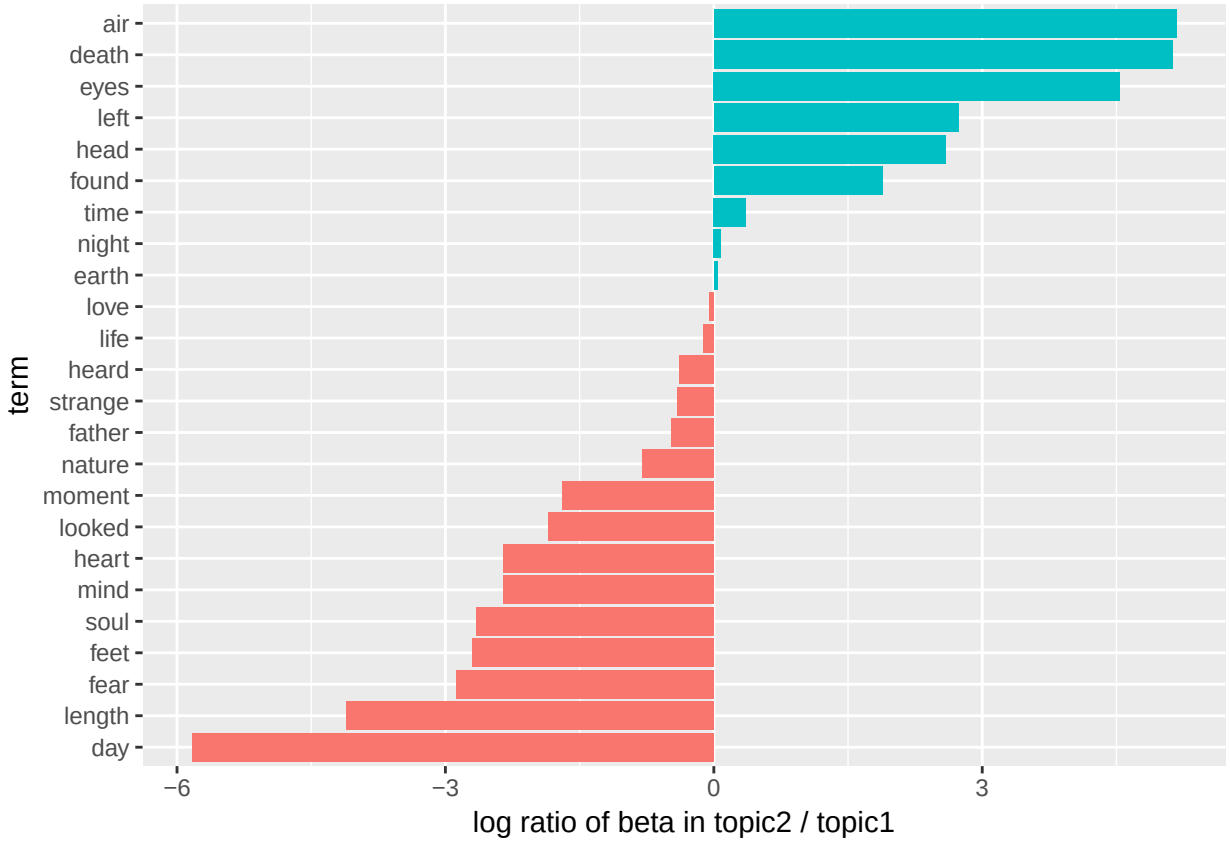
```
spooky_words_lda <- spooky_words_dtm %>%
  LDA(k=4,control=list(seed=1234))

spooky_words_topics <- tidy(spooky_words_lda,matrix="beta")

topic_top_terms <- spooky_words_topics %>%
  group_by(topic) %>%
  top_n(10,beta) %>%
  ungroup %>%
  arrange(topic,-beta)
  # the same: arrange(topic,desc(beta))

topic_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(x=term,y=beta,fill=factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~topic,scales="free") +
  coord_flip()
```



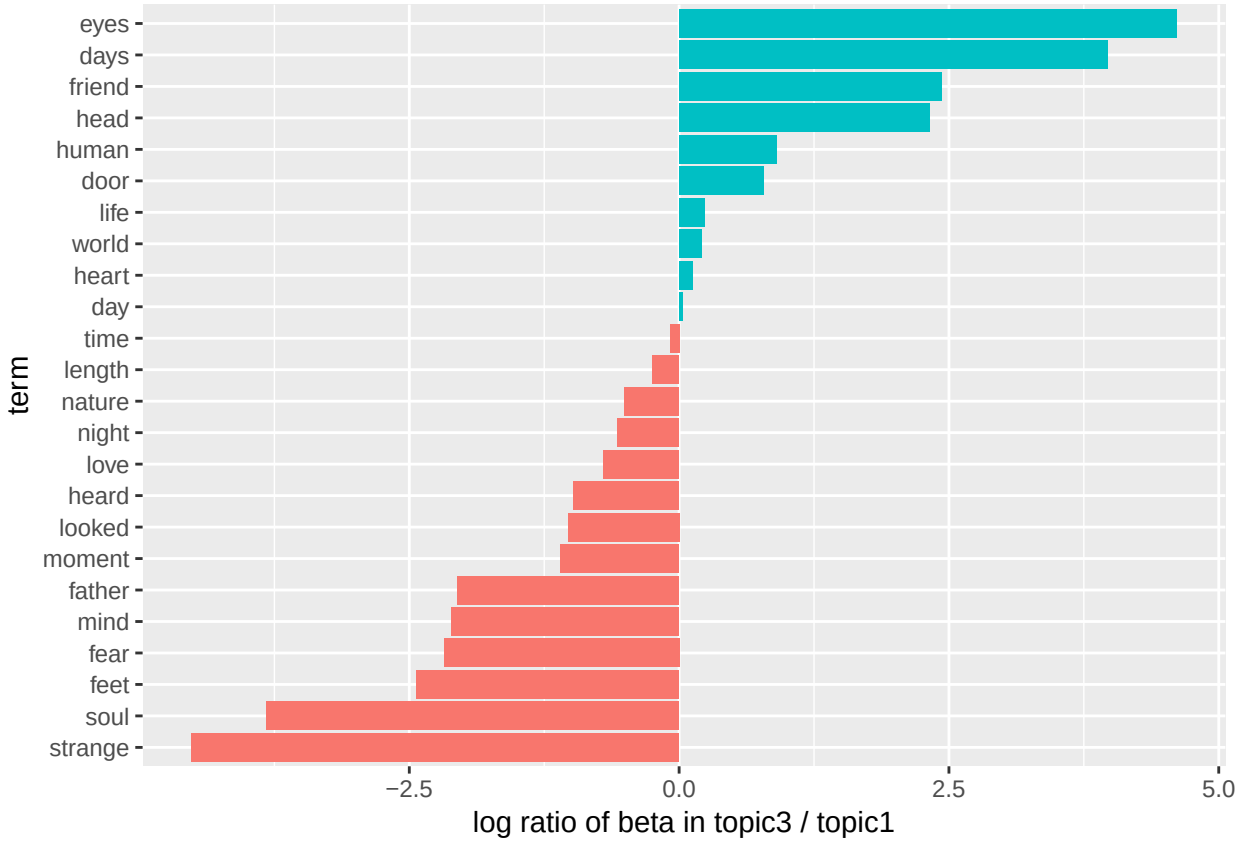
Unlike clustering, it happens that a word could be put under different topics at the same time. We could see that words such as "time", "life", "day", "night", "eyes", "heart" show up in more than two topics here. There's a high proportion of words overlap among topics so it's difficult for me to tell exactly what each topic is about. And becuase the spooky data is extracted from horror stories three authors wrote, it makes sense that the topic doesn't turn out with specific terms like it does with economic news or article about politics. Next step I'll consider the terms that had the greatest difference in beta between the two topics using the log

ratio of the two to discover better differences among the four topics.

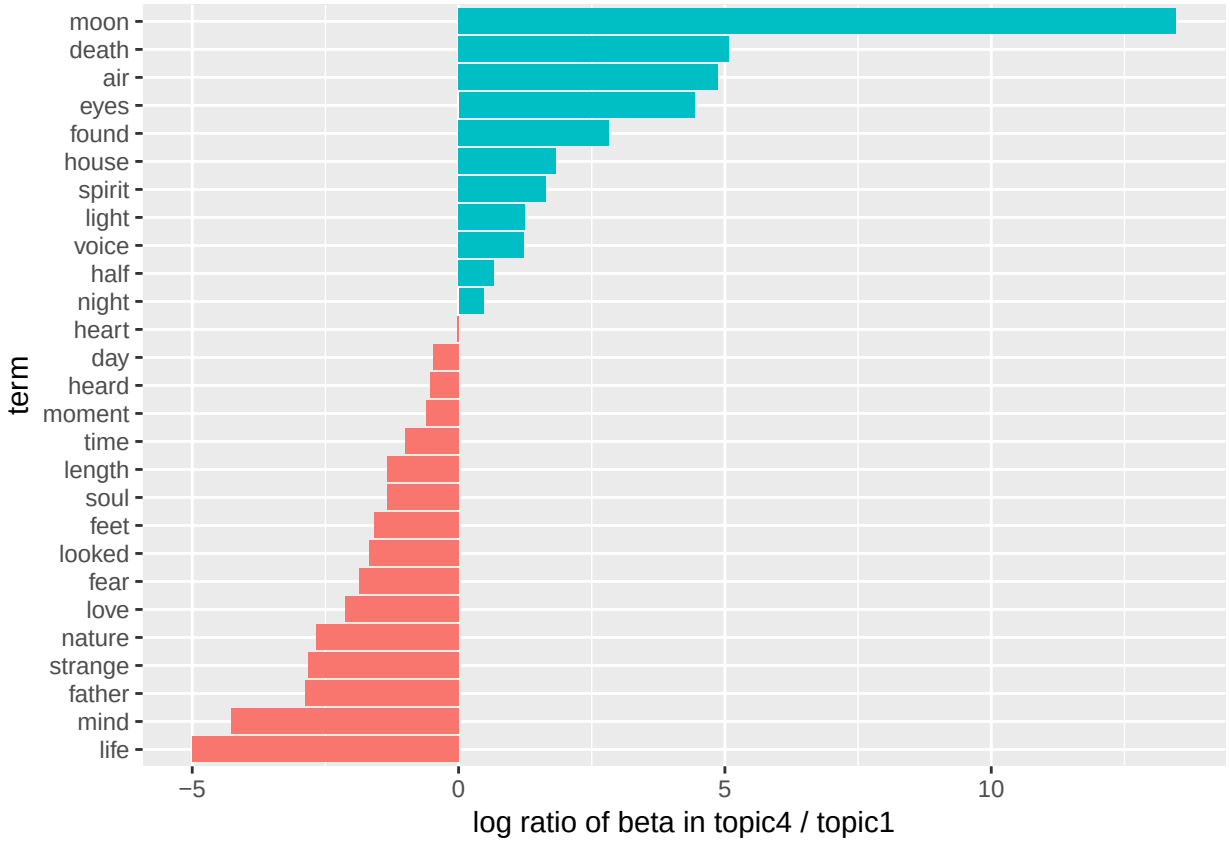
```
beta_spread <- spooky_words_topics %>%
  mutate(topic=paste0("topic",topic)) %>%
  spread(topic,beta)

beta_spread %>%
  filter(topic1>0.002 | topic2>0.002) %>%
  mutate(log_ratio=log2(topic2/topic1)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic2 / topic1") +
  coord_flip()
```



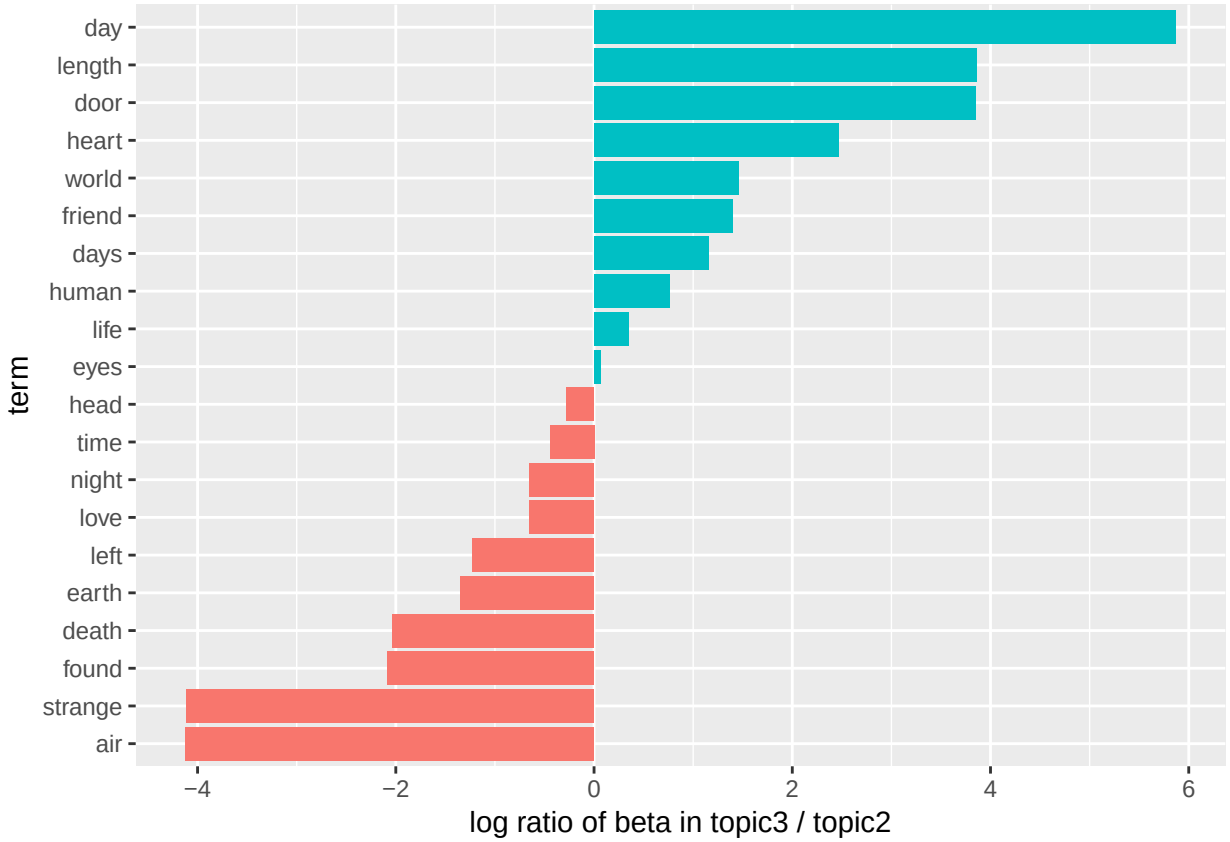

```
beta_spread %>%
  filter(topic1>0.002 | topic3>0.002) %>%
  mutate(log_ratio=log2(topic3/topic1)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic3 / topic1") +
  coord_flip()
```



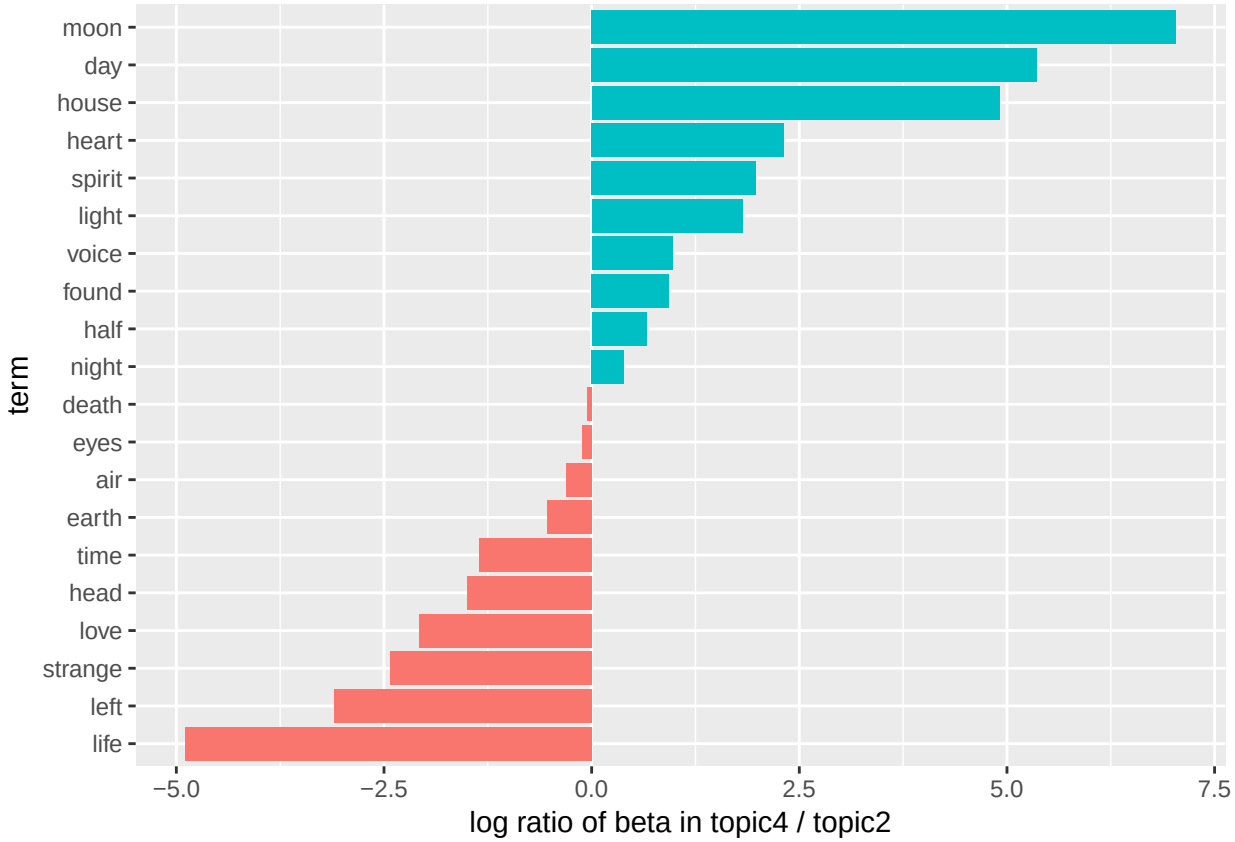

```
beta_spread %>%
  filter(topic1>0.002 | topic4>0.002) %>%
  mutate(log_ratio=log2(topic4/topic1)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic4 / topic1") +
  coord_flip()
```



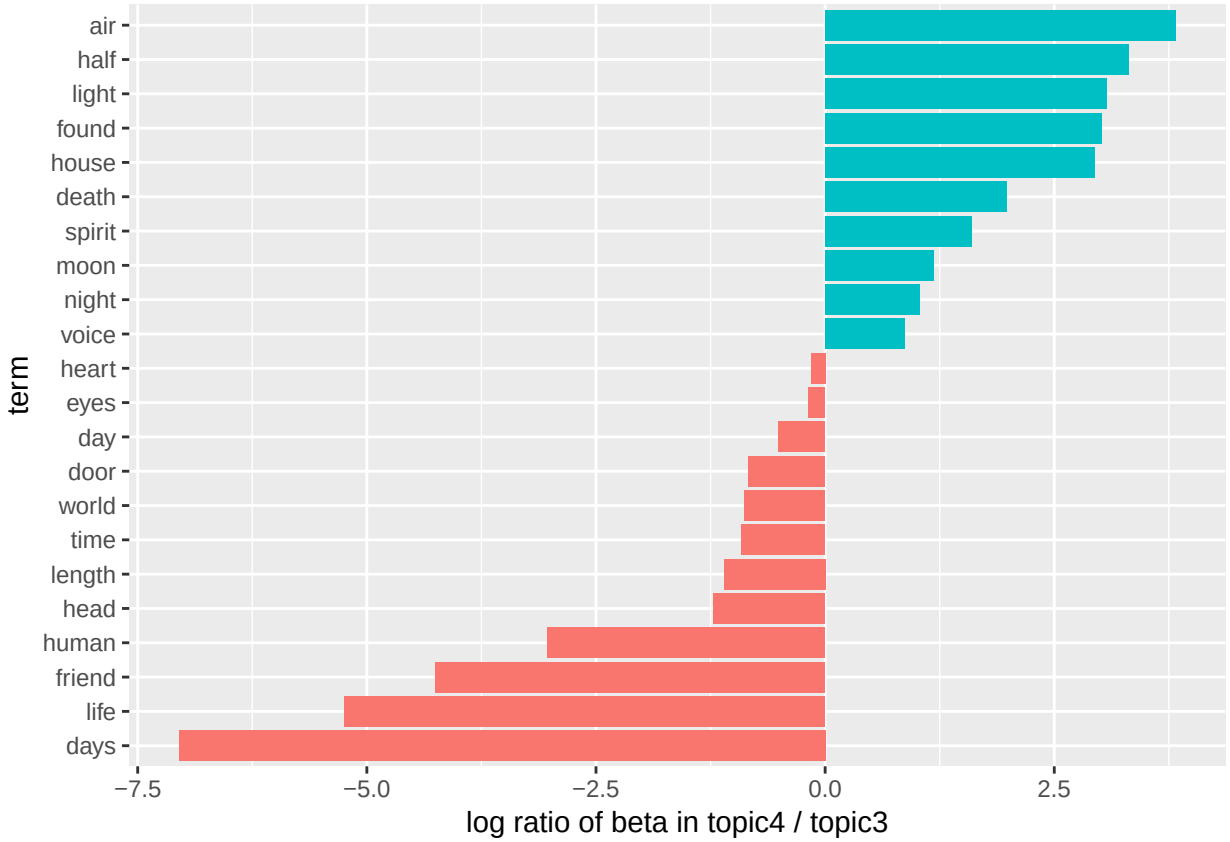

```
beta_spread %>%
  filter(topic2>0.002 | topic3>0.002) %>%
  mutate(log_ratio=log2(topic3/topic2)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic3 / topic2") +
  coord_flip()
```




```
beta_spread %>%
  filter(topic2>0.002 | topic4>0.002) %>%
  mutate(log_ratio=log2(topic4/topic2)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic4 / topic2") +
  coord_flip()
```




```
beta_spread %>%
  filter(topic3>0.002 | topic4>0.002) %>%
  mutate(log_ratio=log2(topic4/topic3)) %>%
  mutate(term=reorder(term,log_ratio)) %>%
  ggplot(aes(x=term,y=log_ratio,fill=log_ratio>0)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(y="log ratio of beta in topic4 / topic3") +
  coord_flip()
```



I found:

- topic4 - mainly contains words that shows depiction of environment and scenes, such as “air”, “light”, “house”, “moon”, night“. It’s more like context before real spooky scenes. It contains”death“”, but not “love”(or beta is < 0.002 so it’s filtered out).
- topic3 - mainly contains words that depicts human feature and characteristics, such as “eyes”, “human”, “friend”, “head”, “human”. It doesn’t contain “death” or “love”.
- topic2 - mixture of things. It has “eyes”, “head”, “life”; has “found”, “left”; has “time”, “night”, “earth”, “air”; has “strange”. It contains both “death” and “love”.
- topic1 - mainly contains words of feelings, such as “fear”, “strange”, “love” and words of action, such as “heard”, “looked”, “moment”. It contains “love”, but not “death”.

Now we’ve understood the mixture of words every topic contain, let’s move on to see how does these topics constitute each author’s work.

Every document is a mixture of topics

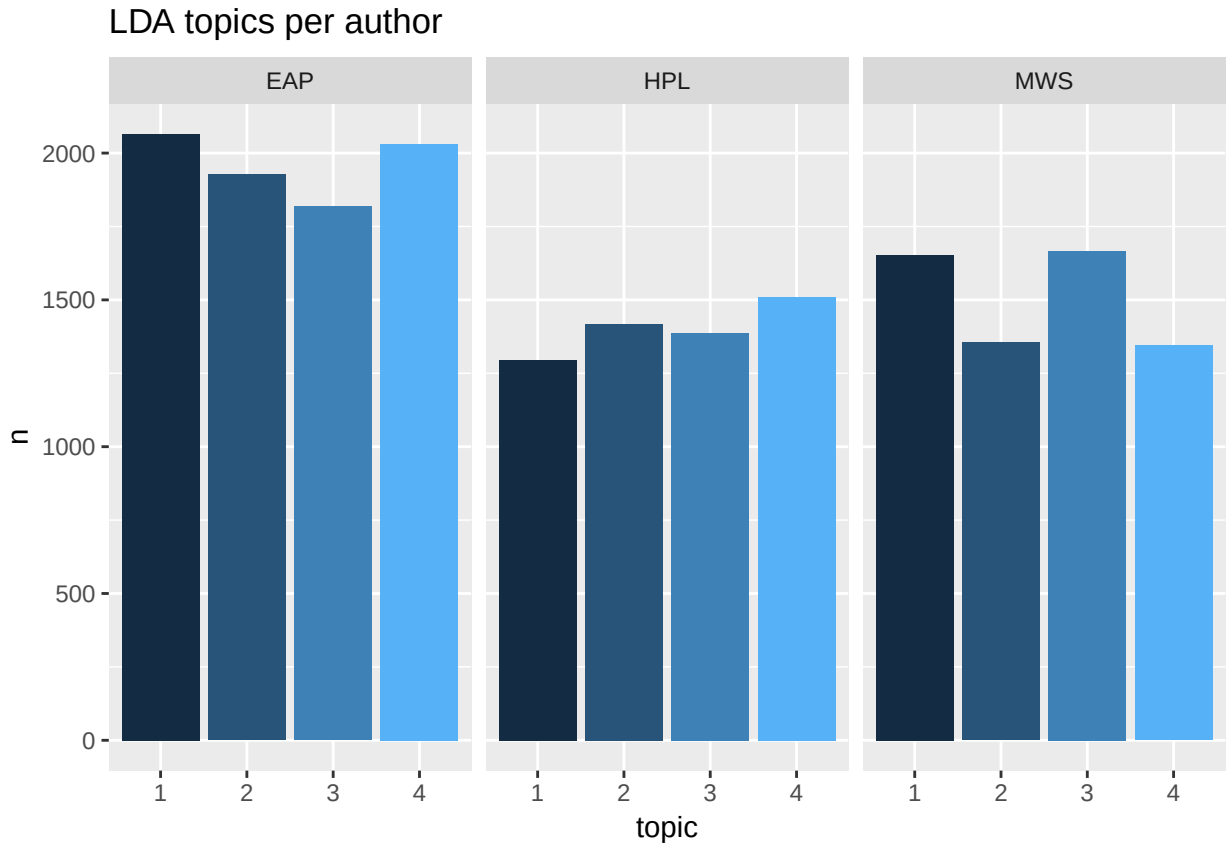
Firstly, I examined the per-sentence-per-topic probabilities:

```
spooky_sentence_topic <- tidy(spooky_words_lda,matrix="gamma")
summary(spooky_sentence_topic$gamma)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1518  0.2473  0.2498  0.2500  0.2524  0.5129
```

It shows 1st quartile, median, mean and 3rd quatile are almost all around 0.25, which means more than half of the sentences in the spooky dataset have an equal probablity (25%) of coming from either topic (1 or 2 or 3 or 4). A higher gamma number that finally assign sentences to only one topic has become not that significant. But that's becuase of the data we have. We still need to see the topic distribution of each author.

```
spooky_author_topic <- spooky_sentence_topic %>%
  left_join(spooky,by=c("document"="id"))
spooky_author_topic$text <- NULL
spooky_author_topic$sen_length <- NULL

# find the topic most associated with each sentence
sentence_classification <- spooky_author_topic %>%
  group_by(author,document) %>%
  top_n(1,gamma) %>%
  ungroup()

# to see topic distribution of each author
sentence_classification %>%
  count(author,topic) %>%
  ggplot(aes(x=factor(topic),y=n,fill=topic)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~author) +
  xlab("topic") +
  ggtitle("LDA topics per author")
```



Though the differences among topics for each author are small, there're still some insights here. Topic 1 and 4 are the top 2 topics for Allen Poe. For Mary Shelly topic 1 and 3 stand out more. Topic 4 is the top topic

for HP Lovecraft and he has a more even distribution of topics than the other two.

There's another plot below to show the same information in a different format. Topics that are on the left side of the bottom are the most significant one in the document. Sequence is in clockwise.

```
spooky_author_topic %>%
  mutate(topic = as.factor(topic)) %>%
  group_by(document) %>%
  top_n(1, gamma) %>%
  ungroup() %>%
  group_by(author, topic) %>%
  count() %>%
  ungroup() %>%
  ggplot(aes(area = n, fill = topic, label = topic, subgroup = author)) +
  geom_treemap() +
  geom_treemap_subgroup_border() +
  geom_treemap_subgroup_text(place = "centre", grow = T, alpha = 0.5, colour =
                               "black", fontface = "italic", min.size = 0) +
  geom_treemap_text(colour = "white", place = "topleft", reflow = T) +
  theme(legend.position = "null") +
  ggtitle("LDA topics per author")
```