Section 4: Topic Models

We use the `topicmodels` package for this analysis. Since the `topicmodels` package doesn't use the `tidytext` framework, we first convert our `spooky_wrd` dataframe into a document term matrix (DTM) matrix using `tidytext` tools.

```
# Counts how many times each word appears in each sentence
sent_wrd_freqs <- count(spooky_wrd, id, word)
head(sent_wrd_freqs)
```

```
## # A tibble: 6 x 3
##   id      word        n
##   <chr>   <chr>   <int>
## 1 id00001 content     1
## 2 id00001 idris       1
## 3 id00001 mine        1
## 4 id00001 of          1
## 5 id00001 resolve     1
## 6 id00001 this        1
```

```
# Creates a DTM matrix
spooky_wrd_tm <- cast_dtm(sent_wrd_freqs, id, word, n)
spooky_wrd_tm
```

```
## <<DocumentTermMatrix (documents: 19579, terms: 25616)>>
## Non-/sparse entries: 444771/501090893
## Sparsity           : 100%
## Maximal term length: 19
## Weighting          : term frequency (tf)
```

```
length(unique(spooky_wrd$id))
```

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

```
length(unique(spooky_wrd$word))
```

```
## [1] 25616
```

The matrix `spooky_wrd_tm` is a sparse matrix with 19467 rows, corresponding to the 19467 ids (or originally, sentences) in the `spooky_wrd` dataframe, and 24941 columns corresponding to the total number of unique words in the `spooky_wrd` dataframe. So each row of `spooky_wrd_tm` corresponds to one of the original sentences. The value of the matrix at a certain position is then the number of occurences of that word (determined by the column) in this specific sentence (determined by the row). Since most sentence/word pairings don't occur, the matrix is sparse meaning there are many zeros.

step1: Determine how many topics to use

Pick 6, 12 # of topics and to see if there are any dupilicate topics or not.

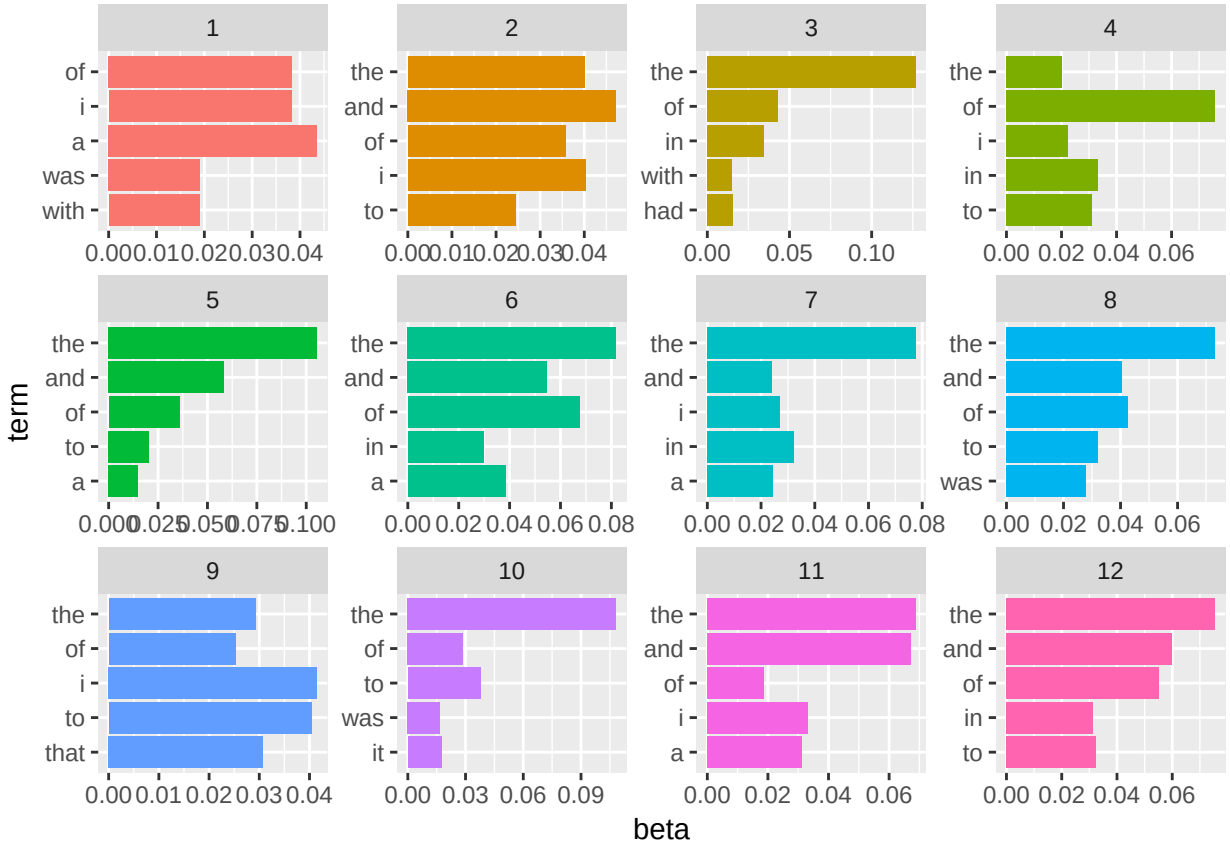
```
spooky_wrd_lda    <- LDA(spooky_wrd_tm, k = 12, control = list(seed = 1234))
spooky_wrd_topics <- tidy(spooky_wrd_lda, matrix = "beta")
spooky_wrd_topics
```

```
## # A tibble: 307,392 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 content 0.0000758
```

```
##  2     2 content 0.0000371
##  3     3 content 0.00000905
##  4     4 content 0.00000642
##  5     5 content 0.0000749
##  6     6 content 0.0000571
##  7     7 content 0.0000414
##  8     8 content 0.0000831
##  9     9 content 0.0000607
## 10    10 content 0.000180
## # ... with 307,382 more rows
```

```
spooky_wrd_topics_5 <- ungroup(top_n(group_by(spooky_wrd_topics, topic), 5, beta))
spooky_wrd_topics_5 <- arrange(spooky_wrd_topics_5, topic, -beta)
spooky_wrd_topics_5 <- mutate(spooky_wrd_topics_5, term = reorder(term, beta))

ggplot(spooky_wrd_topics_5) +
  geom_col(aes(term, beta, fill = factor(topic)), show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
```

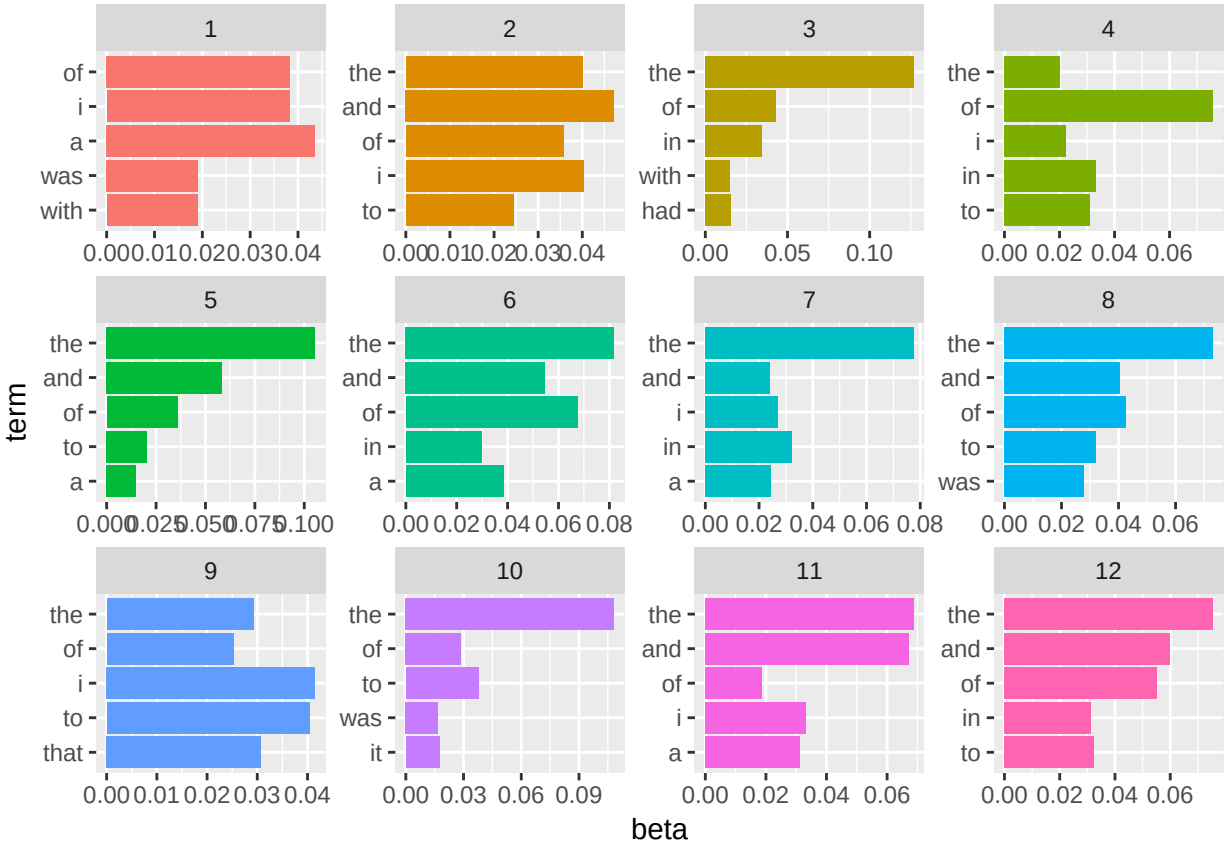



```
spooky_wrd_lda_6<-LDA(spooky_wrd_tm,k=6, control = list(seed = 1234))
spooky_wrd_6_topics <- tidy(spooky_wrd_lda_6, matrix = "beta")
spooky_wrd_6_topics
```

```
## # A tibble: 153,696 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 content 0.000117
```

```
##  2     2 content 0.0000551
##  3     3 content 0.0000141
##  4     4 content 0.00000967
##  5     5 content 0.000110
##  6     6 content 0.0000846
##  7     1 idris   0.000255
##  8     2 idris   0.00000802
##  9     3 idris   0.000300
## 10     4 idris   0.000307
## # ... with 153,686 more rows
```

```
spooky_wrd_6_topics_5 <- ungroup(top_n(group_by(spooky_wrd_topics, topic), 5, beta))
spooky_wrd_6_topics_5 <- arrange(spooky_wrd_topics_5, topic, -beta)
spooky_wrd_6_topics_5 <- mutate(spooky_wrd_topics_5, term = reorder(term, beta))
ggplot(spooky_wrd_6_topics_5) +
  geom_col(aes(term, beta, fill = factor(topic)), show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
```



Compare 6, 12 topic, I would suggest 6, because when u pick 12, there are some kind of of duplicated.

In the above, we see that the first topic is characterized by words like “love”, “earth”, and “words” while the third topic includes the word “thousand”, and the fifth topic the word “beauty”. Note that the words “eyes” and “time” appear in many topics. This is the advantage to topic modelling as opposed to clustering when using natural language – often a word may be likely to appear in documents characterized by multiple topics.