topic modeling

Topic Modeling

```
library(tidyverse)
library(lexicon)
library(factoextra)
library(tidytext)
library(tidygraph)
library(topicmodels)
library(palmerpenguins)
```

Reading in the data:

```
movies <- read.csv("movie_plots.csv")
```

Unnesting tokens using tidytext:

```
plots_by_word <- movies %>% unnest_tokens(word, Plot)
plot_word_counts <- plots_by_word %>%
  anti_join(stop_words) %>%
  count(Movie.Name, word, sort = TRUE)
```

Removing common first names using the 'lexicon package'

```
data("freq_first_names")
first_names <- tolower(freq_first_names$Name)
plot_word_counts <- plot_word_counts %>% filter(!(word %in% first_names))
```

Casting our word counts to a document term matrix

```
plots_dtm <- plot_word_counts %>% cast_dtm(Movie.Name, word, n)
```

Before LDA a look at the dimensions of our matrix:

```
#Distinct words
dim(plot_word_counts %>% distinct(word))[1]
```

```
[1] 13394
```

```
dim(movies)
```

```
[1] 1077    2
```

LDA with 30 topics

```
plots_lda <- LDA(plots_dtm, k = 25, control = list(seed = 1066))
```

Retrieving gammas:

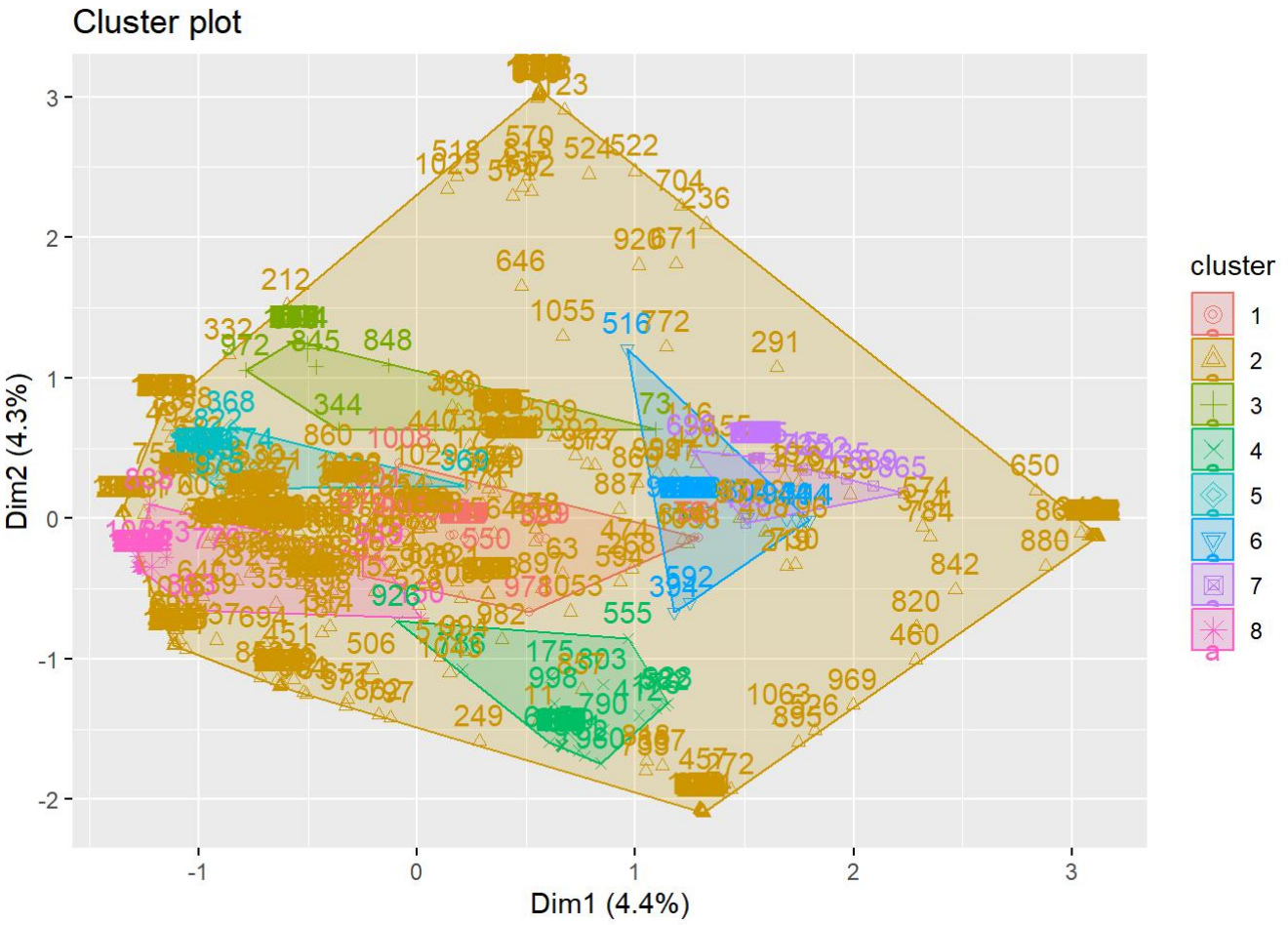
```
plots_gamma <- tidy(plots_lda, matrix = "gamma")
```

Pivoting the plots_gamma table wider so we can cluster by gammas for each topic

```
plots_gamma_wider <- plots_gamma %>% pivot_wider(
  names_from = topic,
  values_from = gamma
)
```

Clustering. We are considering 10 genres. So lets try 10 clusters

```
plots_gamma_wider_no_na <- plots_gamma_wider %>% drop_na()
cluster <- kmeans(plots_gamma_wider %>% select(-document), 8)
fviz_cluster(cluster, data = plots_gamma_wider %>% select(-document))
```



Let's look at the genres in each cluster: So we'll read in the data with the genres

```
english_movies_with_genres <- read.csv("movie_plots_with_genres.csv")
clusters <- cluster[["cluster"]]
plots_gamma_wider$cluster <- clusters
```

Cluster 7:

```r
plots_clusters7 <- plots_gamma_wider %>% filter(cluster == 5)
cluster_7_names <- plots_clusters7$document
cluster_7 <- english_movies_with_genres %>% filter(Movie.Name %in% cluster_7_names)
cluster_7_counts <- cluster_7 %>% group_by(Genre) %>% summarize(n = n())
```