Untitled

2024-11-09

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
movies<-read.csv("movie_plots_with_genres.csv")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(janeaustenr)
library(tidytext)
library(topicmodels)
library(tidyr)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
data("stop_words")
movie_words <- movies |> unnest_tokens(word, Plot)
movie_counts <- movie_words %>%
  anti_join(stop_words) %>%
  count(Movie.Name, word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

Weeding out the names, reorganize the data:

```
library(lexicon)
data("freq_first_names")
firstname <- tolower(freq_first_names$Name)
movie_counts <- movie_counts |> filter(!(word %in% firstname))
```

Casting the words counts to a matrix

```
counts_matrix<-movie_counts |> cast_dtm(Movie.Name,word,n)
```

```
example <- head(counts_matrix, n=6)
print(example)
```

```
## <<DocumentTermMatrix (documents: 6, terms: 13396)>>
## Non-/sparse entries: 638/79738
## Sparsity           : 99%
## Maximal term length: 17
## Weighting          : term frequency (tf)
```

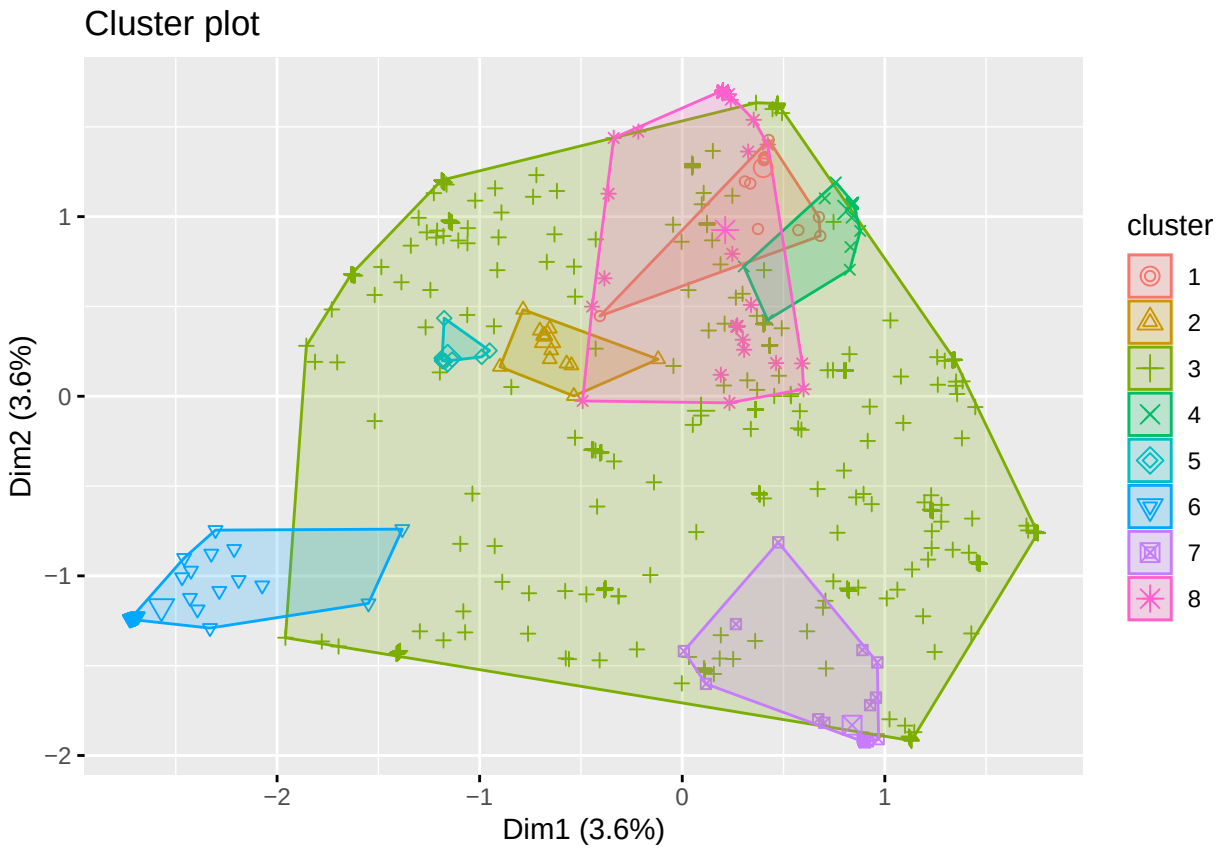
```
dim(movie_counts)
```

```
## [1] 44142     3
```

```
dim(movies)
```

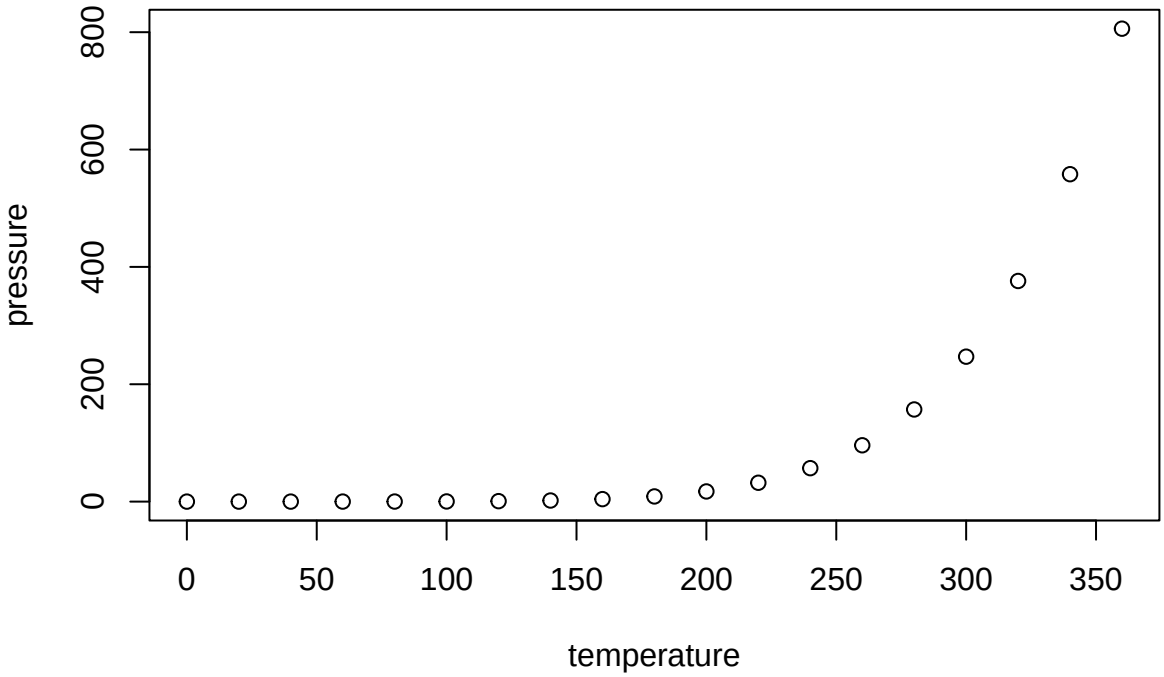
```
## [1] 1077    4
```

```
lda <- LDA(counts_matrix, k = 30, control = list(seed = 1066))
plots_gamma <- tidy(lda, matrix = "gamma") %>%
  pivot_wider(names_from = topic, values_from = gamma) %>%
  drop_na()
cluster <- kmeans(select(plots_gamma, -document), centers = 8, nstart = 25)
fviz_cluster(cluster, data = select(plots_gamma, -document), geom = "point")
```



Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.