Web scraping first time

IMBD

```
library("rvest")
```

```
## Loading required package: xml2
```

```
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```
library("stringr")
library("tidytext")
library("glmnet")
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.0
```

```
theme_set(theme_minimal())

read_webpg <- read_html("https://www.imdb.com/list/ls041125816/")
```

Titre

```
titre <- read_webpg %>%
         html_nodes(".lister-item-header")%>%
         html_text() %>%
         str_replace_all(.,"[:digit:]","")%>%
         str_replace_all(.,"\\\n","")%>%
         str_replace_all("[:punct:]","") %>%
         str_replace_all(.,"\\s\\I","") %>%
         str_trim(side = "both")

titre[17] <- '1917'
length(titre)
```

```
## [1] 100
```

Genre

```
genre <- read_webpg %>%
         html_nodes(".genre")%>%
         html_text()%>%
         str_replace_all(.,"[\r\n]","")%>%
         str_replace_all(.,"\\s","")


length(genre)
```

```
## [1] 100
```

Score

```
read_webpg %>%
  html_nodes("span.ipl-rating-star__rating")%>%
  html_text() -> rate


rate_nb <- rate[seq(1,length(rate),23)]

length(rate_nb)
```

```
## [1] 100
```

Directors

```
directors <- read_webpg %>%
  html_nodes('.text-small a:nth-child(1)' )%>%
  html_text()
```

```
directors  <- directors[-1]

length(directors)
```

```
## [1] 100
```

Stars

```
Principal_stars <- read_webpg %>%
                   html_nodes(".text-small:nth-child(5) a:nth-child(1) , .ghost+ a") %>%
                   html_text()

length(Principal_stars)
```

```
## [1] 100
```

Review

.ipl-rating-widget+ p , .ratings-metascore+ p

```
review <- read_webpg %>%
  html_nodes(".ipl-rating-widget+ p , .ratings-metascore+ p")%>%
  html_text()%>%
  str_replace_all(.,"\n","")%>%
  str_trim()

length(review)
```

```
## [1] 100
```

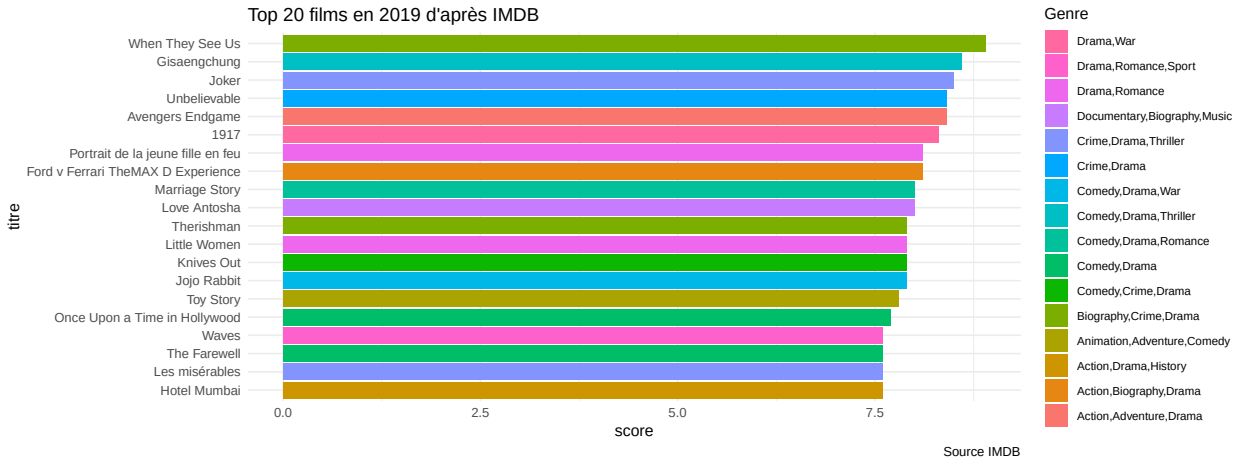
to Fix , #add director # add description

Data frame

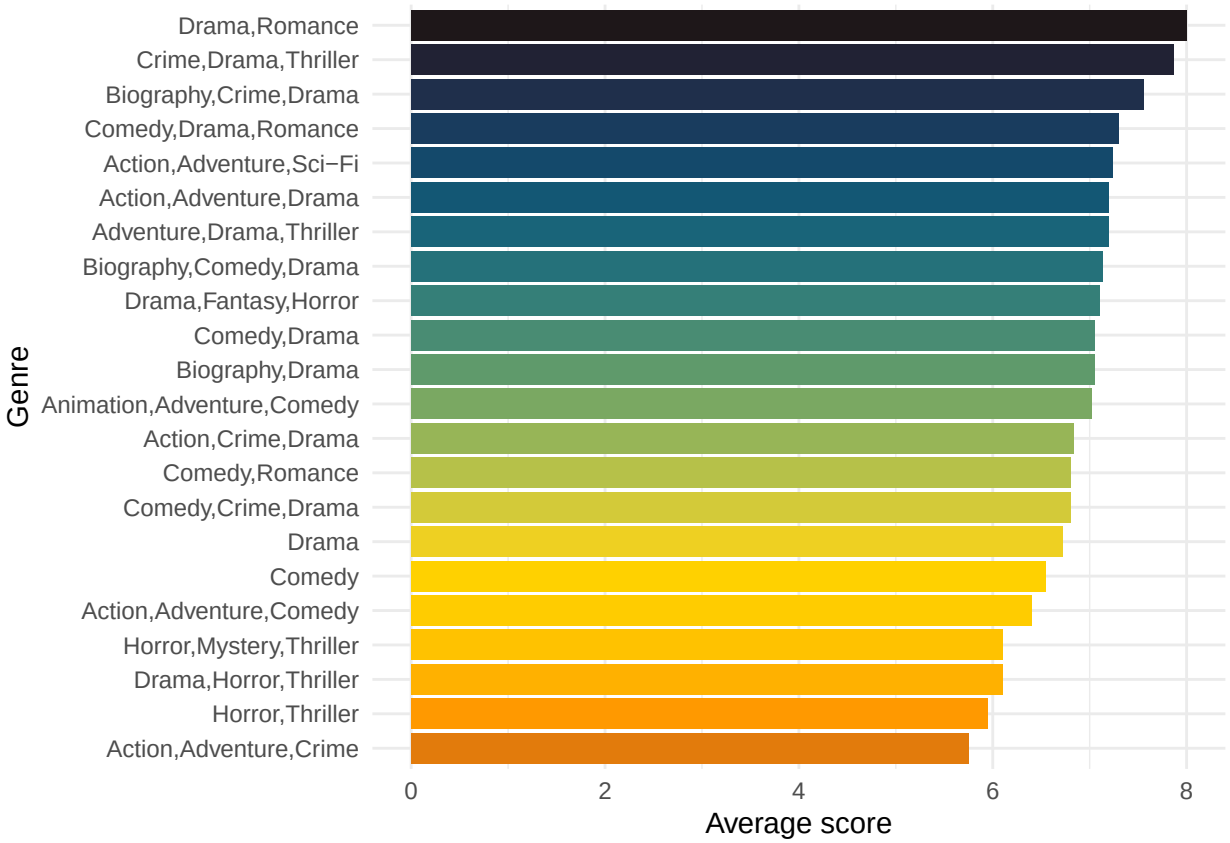
```
df <- data.frame(titre = titre , genre = genre , score = rate_nb,directors = directors , actor_1 = Prin
row.names(df) <- NULL

view(df)
```

```
df %>%
  select(-review)%>%
  arrange(desc(score))%>%
  head(20)%>%
  mutate(score = as.numeric(score),
         titre= fct_reorder(titre,score))%>%
  ggplot(mapping = aes(x = titre , y = score , fill = as.factor(genre)))+
```

```
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  guides(fill = guide_legend( reverse = TRUE))+
  labs(fill = "Genre" , title = "Top 20 films en 2019 d'après IMDB", caption = "Source IMDB")+
  theme(legend.text = element_text(size = 8))
```




```
library(fishualize)

df %>%
  group_by(genre)%>%
  mutate(nb = n(),
         score = as.numeric(score))%>%
  filter(nb > 1 ) %>%
  summarise(avg_score = mean(score))%>%
  mutate(genre = fct_reorder(genre,avg_score))%>%
  ggplot(mapping = aes(x = genre , y = avg_score,fill = genre))+
  geom_bar(stat = "identity")+
  scale_fill_fish_d(option = "Balistapus_undulatus")+
  coord_flip() +  guides(fill = guide_legend( reverse = TRUE))+
  theme(legend.position = "")+
  labs(y = "Average score" , x = "Genre" )
```



Next time add comment and do a lasso regression to check how words (theme of moovie) influence the rating of the movie

Analysing actors & Directors

```
df %>%
  count(directors,sort = TRUE) #Only Jeff Chan directed 2 moovies in the top 100
```

```
## # A tibble: 98 x 2
##    directors               n
##    <chr>               <int>
##  1 Jeff Chan               2
##  2 Steven Soderbergh       2
##  3 Abe Forsythe            1
##  4 Adam Egypt Mortimer     1
##  5 Adam Robitel            1
##  6 Alejandro Landes        1
##  7 Alex Lehmann            1
##  8 Alma Har'el             1
##  9 Andy Muschietti         1
## 10 Anna Boden              1
## # ... with 88 more rows
```

```
liste_actors <- df %>%
  count(actor_1,sort = TRUE)%>%
  filter(n >1)%>%
  pull(actor_1)
```

```
df %>%
  filter(actor_1 %in% liste_actors)%>%
  view()
```

```
features_1 <- df %>%
  unnest_tokens(word,review)%>%
  anti_join(stop_words,by = "word")%>%
  add_count(word) %>%
  separate_rows(genre  ,sep  =",")%>%
  mutate(genre_row = paste0(genre,":",word))%>%
  select(-score) %>%
    mutate(value = 1)


feat_mat <- features_1 %>%
  cast_sparse(titre,genre_row)

feat_mat
```

```
## 100 x 3083 sparse Matrix of class "dgCMatrix"

##    [[ suppressing 33 column names 'Comedy:greed', 'Drama:greed', 'Thriller:greed' ... ]]

##
## Gisaengchung                          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## Uncut Gems                            . . . . . . . . . . . . . . . . . . .
## Portrait de la jeune fille en feu     . . . . . . . . . . . . . . . . . . .
## Midsommar                             . . . . . . . . . . . . . . . . . . .
## Joker                                 . . . . . . . . . . . . . . . . . . .
## The Farewell                          . . . . . . . . . . . . . . . . . . .
## Hatsukoi                              . . . . . . . . . . . . . . . . . . .
## Waves                                 . . . . . . . . . . . . . . . . . . .
## Les misérables                        . . . . . . . . . . . . . . . . . . .
## Ford v Ferrari TheMAX D Experience    . . . . . . . . . . . . . . . . . . .
## Marriage Story                        . . . . . . . . . . . . . . . . . . .
## Hotel Mumbai                          . . . . . . . . . . . . . . . . . . .
## El Camino A Breaking Bad Movie        . . . . . . . . . . . . . . . . . . .
## The Peanut Butter Falcon              . . . . . . . . . . . . . . . . . . .
## Jai perdu mon corps                   . . . . . . . . . . . . . . . . . . .
## Therishman                            . . . . . . . . . . . . . . . . . . .
## 1917                                  . . . . . . . . . . . . . . . . . . .
## The Art of SelfDefense                . . . . . . . . . . . . . . . . . . .
## The Nightingale                       . . . . . . . . . . . . . . . . . . .
## Shazam                                . . . . . . . . . . . . . . 1 . . . . .
## Honey Boy                             . . . . . . . . . . . . . . . . . . .
## Dolemites My Name                     . . . . . . . . . . . . . . . . . . .
## Arctic                                . . . . . . . . . . . . . . . . . . .
## The Lighthouse                        . . . . . . . . . . . . . . . . . . .
## Love Antosha                          . . . . . . . . . . . . . . . . . . .
## Richard Jewell                        . . . . . . . . . . . . . . . . . . .
## Triple Frontier                       . . . . . . . . . . . . . . . . . . .
```

```
## Luce                                    . . . . . .
## Little Monsters                         . . . . . .
## The King                                . . . . . .
## Monos                                   . . . . . .
## Us   I                                  . . . . . .
## Come to Daddy                           . . . . . .
## Knives Out                              . . . . . .
## Villains                                . . . . . .
## SpiderMan Far from Home                 . . . . . .
## Avengers Endgame                        . . . . . .
## Am Mother                               . . . . . .
## High Flying Bird                        . . . . . .
## Booksmart                               . . . . . .
## Fighting with My Family                 . . . . . .
## John Wick Chapter   Parabellum          . . . . . .
## Godzilla King of the Monsters           . . . . . .
## Always Be My Maybe                      . . . . . .
## The Laundromat                          . . . . . .
## Dragged Across Concrete                 . . . . . .
## The Perfection                          . . . . . .
## Extremely Wicked Shockingly Evil and Vile . . . . . .
## Plus One                                . . . . . .
## Fyre                                    . . . . . .
## The Breaker Upperers                    . . . . . .
## Color Out of Space                      . . . . . .
## Guns Akimbo                             . . . . . 1 . . . . .
## Danielsnt Real                          . . . . . .
## Late Night                              . . . . . .
## Jojo Rabbit                             . . . . . .
## Once Upon a Time in Hollywood           . . . . . .
## The Lodge                               . . . . . .
## Escape Room                             . . . . . .
## Bombshell                               . . . . . .
## Code                                    . . . . . .
## The Lego Movie  The Second Part         . . . . . .
## A Shaun the Sheep Movie Farmageddon     . . . . . .
## See You Yesterday                       . . . . . .
## Captain Marvel                          . . . . . .
## A Vigilante                             . . . . . .
## Hustlers                                . . . . . .
## Long Shot                               . . . . . .
## Toy Story                               . . . . . .
## Angel Has Fallen                        . . . . . .
## t Chapter Two                           . . . . . .
## Bloodline   II                          . . . . . .
## Brittany Runs a Marathon                . . . . . .
## The Dirt                                . . . . . .
## FrozenI                                 . . . . . .
## The Highwaymen                          . . . . . .
## Lady and the Tramp                      . . . . . .
## American Son                            . . . . . .
## Happy Death Day U                       . . . . . .
## sntt Romantic                           . . . . . .
## Paddleton                               . . . . . .
```

```
## The Boy Who Harnessed the Wind           . . . . . .
## Glass                                    . . . . . .
## Alita Battle Angel                       . . . . . .
## The Mustang                              . . . . . .
## Captive State                            . . . . . .
## The Curse of la Llorona                  . . . . . .
## The Prodigy                              . . . . . .
## The Kid Who Would Be King                . . . . . .
## The Secret Life of Pets                  . . . . . .
## The Goldfinch                            . . . . . .
## Ad Astra                                 . . . . . .
## Men in Blacknternational                 . . . . . .
## Rocketman                                . . . . . .
## Little Women                             . . . . . .
## Pokémon Detective Pikachu                . . . . . .
## Whered You Go Bernadette                 . . . . . .
## A Beautiful Day in the Neighborhood      . . . . . .
## Unbelievable                             . . . . . .
## When They See Us                         . . . . . .
##
## Gisaengchung                             1 1 1 1 1 1 1 1 1 1 1 1 1 1 ......
## Uncut Gems                               . . . . . .
## Portrait de la jeune fille en feu        . . . . . .
## Midsommar                                . . . . . .
## Joker                                    . . . . . .
## The Farewell                             . . . . . 1 1 . ......
## Hatsukoi                                 . . . . . .
## Waves                                    . . . . . 1 . ......
## Les misérables                           . . . . . .
## Ford v Ferrari TheMAX D Experience       . . . . . .
## Marriage Story                           . . . . . 1 1 . ......
## Hotel Mumbai                             . . . . . .
## El Camino A Breaking Bad Movie           . . . . . .
## The Peanut Butter Falcon                 . . . . . .
## Jai perdu mon corps                      . . . . . .
## Therishman                               . . . . . .
## 1917                                     . . . . . .
## The Art of SelfDefense                   . . . . . .
## The Nightingale                          . . . . . 1 1 ......
## Shazam                                   . . . . . .
## Honey Boy                                . . . . . .
## Dolemites My Name                        . . . . . .
## Arctic                                   . . . . . .
## The Lighthouse                           . . . . . .
## Love Antosha                             . . . . . .
## Richard Jewell                           . . . . . .
## Triple Frontier                          . . . . . .
## Luce                                     . . . . . .
## Little Monsters                          . . . . . .
## The King                                 . . . . . .
## Monos                                    . . . . . .
## Us   I                                   . . . . . .
## Come to Daddy                            . . . . . .
## Knives Out                               . . . . . 1 1 . ......
```

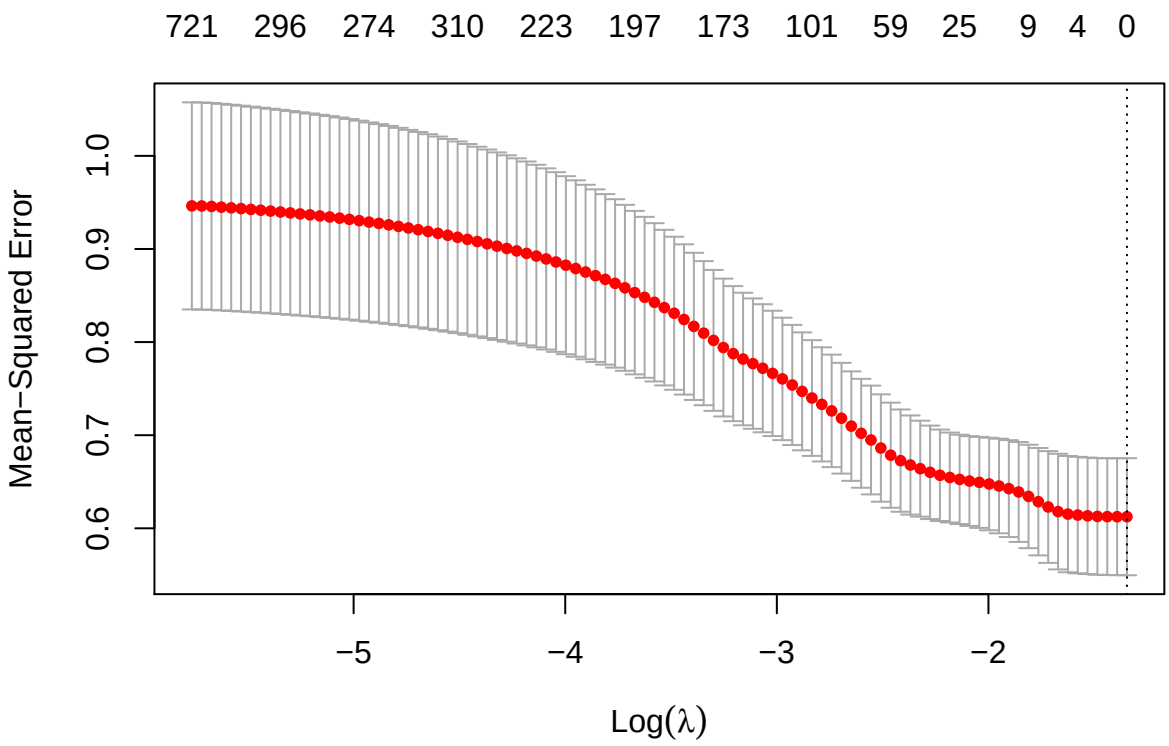
```
## Villains                                  . . . . . .
## SpiderMan Far from Home                   . . . . . .
## Avengers Endgame                          . . . . . .
## Am Mother                                 . . . . . .
## High Flying Bird                          . . . . . .
## Booksmart                                 . . . . . .
## Fighting with My Family                   . . . . . 1 1 . . . . . .
## John Wick Chapter   Parabellum            . . . . . .
## Godzilla King of the Monsters             . . . . . .
## Always Be My Maybe                        . . . . . .
## The Laundromat                            . . . . . .
## Dragged Across Concrete                   . . . . . .
## The Perfection                            . . . . . .
## Extremely Wicked Shockingly Evil and Vile . . . . . .
## Plus One                                  . . . . . .
## Fyre                                      . . . . . .
## The Breaker Upperers                      . . 1 . . . . . .
## Color Out of Space                        . . . . . .
## Guns Akimbo                               . . . . . .
## Danielsnt Real                            . . . . . 1 . . . . . .
## Late Night                                . . . . . .
## Jojo Rabbit                               . . . . . .
## Once Upon a Time in Hollywood             . . . . . .
## The Lodge                                 . . . . . .
## Escape Room                               . . . . . .
## Bombshell                                 . . . . . .
## Code                                      . . . . . .
## The Lego Movie  The Second Part           . . . . . .
## A Shaun the Sheep Movie Farmageddon       . . . . . .
## See You Yesterday                         . . . . . .
## Captain Marvel                            . . . . . .
## A Vigilante                               . . . . . .
## Hustlers                                  . . . . . .
## Long Shot                                 . . . . . .
## Toy Story                                 . . . . . .
## Angel Has Fallen                          . . . . . .
## t Chapter Two                             . . . . . .
## Bloodline   II                            . . . . . 1 . . . . . .
## Brittany Runs a Marathon                  . . . . . .
## The Dirt                                  . . . . . .
## FrozenI                                   . . . . . .
## The Highwaymen                            . . . . . .
## Lady and the Tramp                        . . . . . .
## American Son                              . . . . . .
## Happy Death Day U                         . . . . . .
## sntt Romantic                             . . . . . .
## Paddleton                                 . . . . . .
## The Boy Who Harnessed the Wind            . . . . . 1 . . . . . .
## Glass                                     . . . . . .
## Alita Battle Angel                        . . . . . .
## The Mustang                               . . . . . .
## Captive State                             . . . . . .
## The Curse of la Llorona                   . . . . . .
## The Prodigy                               . . . . . .
```

```
## The Kid Who Would Be King                . . . . . .
## The Secret Life of Pets                  . . . . . .
## The Goldfinch                            . . . . . .
## Ad Astra                                 . . . . . .
## Men in Blacknternational                 . . . . . .
## Rocketman                                . . . . . .
## Little Women                             . . . . . .
## Pokémon Detective Pikachu                . . . . . .
## Whered You Go Bernadette                 . . . . . .
## A Beautiful Day in the Neighborhood      . . . . . .
## Unbelievable                             . . . . . .
## When They See Us                         . . . . . .
##
##  .....suppressing 3050 columns in show(); maybe adjust 'options(max.print= *, width = *)'
##  ..............................
```

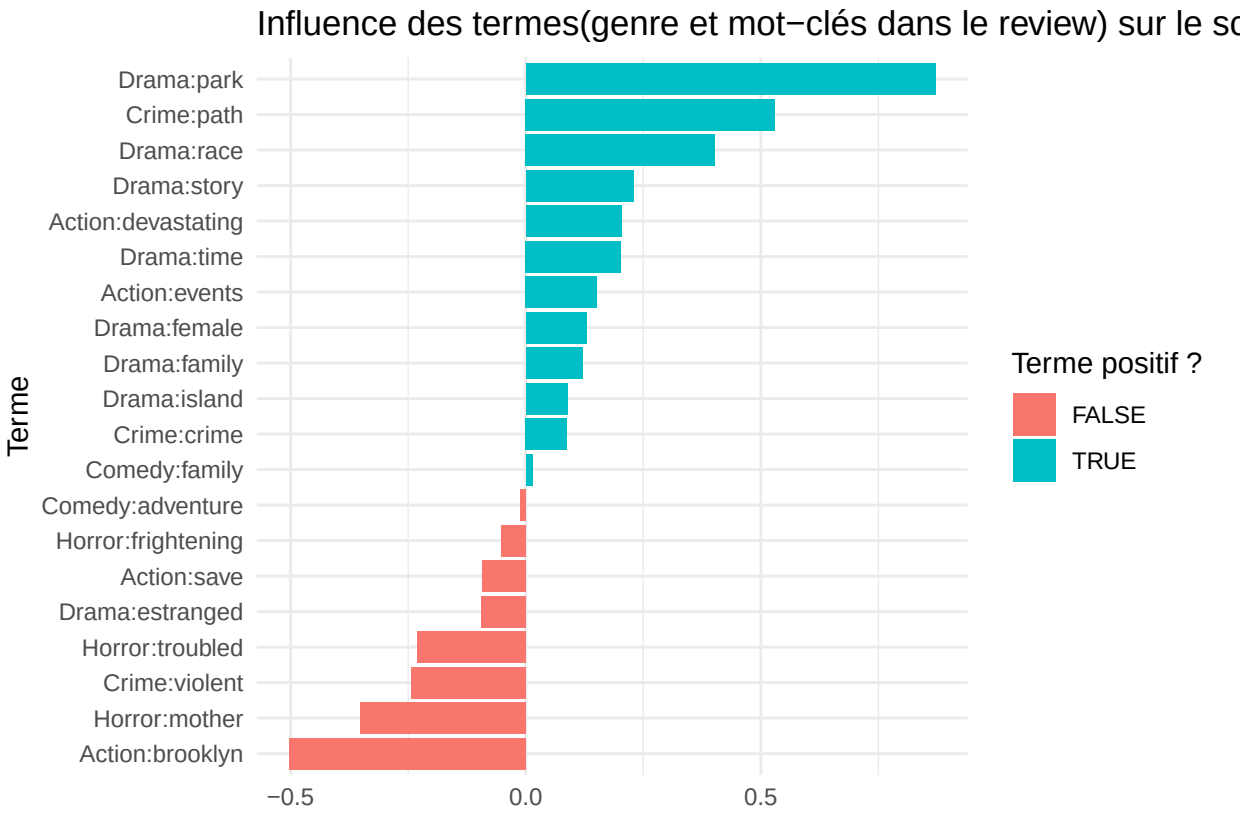
```
dim(feat_mat)
```

```
## [1]  100 3083
```

```
df$score <- as.numeric(df$score)

score <- df$score[match(rownames(feat_mat),df$titre)]

modele <- cv.glmnet(feat_mat,score)

plot(modele)
```




```
tidy(modele$glmnet.fit) %>%
  mutate(lambda  = round(lambda,3)) %>%
  filter(lambda == 0.118, term != "(Intercept)", abs(round(estimate , 3)) > 0.005)%>%
  mutate(term = fct_reorder(term,estimate))%>%
  ggplot(mapping = aes(x = term , y = estimate, fill = estimate >0 ))+
  geom_col()+
  coord_flip()+
  labs(x = "Terme", y = "" , title = "Influence des termes(genre et mot-clés dans le review) sur le sco
```


Influence des termes(genre et mot−clés dans le review) sur le sc
Terme
Drama:park
Crime:path
Drama:race
Drama:story
Action:devastating
Drama:time
Action:events
Drama:female
Drama:family
Drama:island
Crime:crime
Comedy:family
Comedy:adventure
Horror:frightening
Action:save
Drama:estranged
Horror:troubled
Crime:violent
Horror:mother
Action:brooklyn
−0.5
0.0
0.5
Terme positif ?
FALSE
TRUE