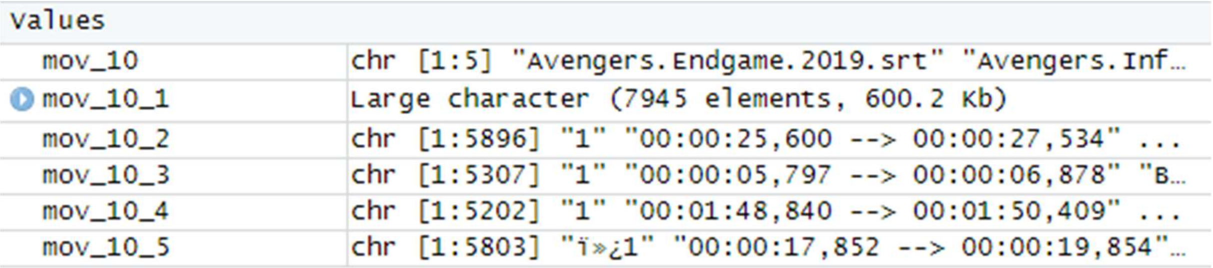
R code | Output

```
# Loading packages: ####
library(tidyverse)
library(textreadr)
library(tidytext)
library(wordcloud)
library(RColorBrewer)
library(reshape2)
library(plotly)

# Importing top 5 highest revenues per decade: ####

## 2010's:
setwd("C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/2010")
mov_10 <- list.files(path="C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/2010")
# Using read document to import the data:
mov_10_1 <- read_document(file=mov_10[1])
mov_10_2 <- read_document(file=mov_10[2])
mov_10_3 <- read_document(file=mov_10[3])
mov_10_4 <- read_document(file=mov_10[4])
mov_10_5 <- read_document(file=mov_10[5])

## 2000's:
setwd("C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/2000")
mov_00 <- list.files(path="C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/2000")
# Using read document to import the data:
mov_00_1 <- read_document(file=mov_00[1])
mov_00_2 <- read_document(file=mov_00[2])
mov_00_3 <- read_document(file=mov_00[3])
mov_00_4 <- read_document(file=mov_00[4])
mov_00_5 <- read_document(file=mov_00[5])

## 1990's:
setwd("C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/1990")
mov_90 <- list.files(path="C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/1990")
# Using read document to import the data:
mov_90_1 <- read_document(file=mov_90[1])
mov_90_2 <- read_document(file=mov_90[2])
mov_90_3 <- read_document(file=mov_90[3])
mov_90_4 <- read_document(file=mov_90[4])
```

```
mov_90_5 <- read_document(file=mov_90[5])

## 1980's:
setwd("C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/1980")
mov_80 <- list.files(path="C:/Users/thiag/Thiago/Hult/Text Analytics/Individual Assignment/1980")
# Using read document to import the data:
mov_80_1 <- read_document(file=mov_80[1])
mov_80_2 <- read_document(file=mov_80[2])
mov_80_3 <- read_document(file=mov_80[3])
mov_80_4 <- read_document(file=mov_80[4])
mov_80_5 <- read_document(file=mov_80[5])

# Example:
```

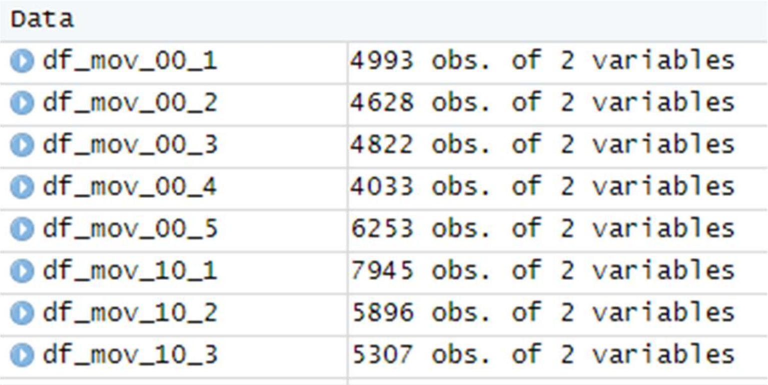


```
Values
  mov_10      chr [1:5] "Avengers.Endgame.2019.srt" "Avengers.Inf…
  mov_10_1    Large character (7945 elements, 600.2 Kb)
  mov_10_2    chr [1:5896] "1" "00:00:25,600 --> 00:00:27,534" ...
  mov_10_3    chr [1:5307] "1" "00:00:05,797 --> 00:00:06,878" "B…
  mov_10_4    chr [1:5202] "1" "00:01:48,840 --> 00:01:50,409" ...
  mov_10_5    chr [1:5803] "ï»¿1" "00:00:17,852 --> 00:00:19,854"…
```

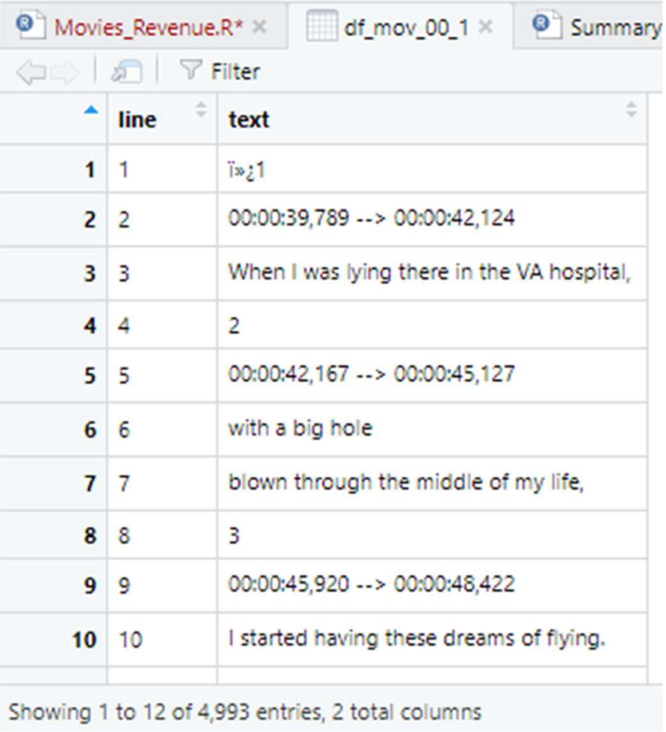
```
# Converting each movie in a dataframe: ####

## 2010's:
df_mov_10_1 <- data_frame(line=1:7945, text=mov_10_1)
df_mov_10_2 <- data_frame(line=1:5896, text=mov_10_2)
df_mov_10_3 <- data_frame(line=1:5307, text=mov_10_3)
df_mov_10_4 <- data_frame(line=1:5202, text=mov_10_4)
df_mov_10_5 <- data_frame(line=1:5803, text=mov_10_5)
## 2000's:
df_mov_00_1 <- data_frame(line=1:4993, text=mov_00_1)
df_mov_00_2 <- data_frame(line=1:4628, text=mov_00_2)
df_mov_00_3 <- data_frame(line=1:4822, text=mov_00_3)
df_mov_00_4 <- data_frame(line=1:4033, text=mov_00_4)
df_mov_00_5 <- data_frame(line=1:6253, text=mov_00_5)
## 1990's:
df_mov_90_1 <- data_frame(line=1:7487, text=mov_90_1)
df_mov_90_2 <- data_frame(line=1:3864, text=mov_90_2)
df_mov_90_3 <- data_frame(line=1:5352, text=mov_90_3)
df_mov_90_4 <- data_frame(line=1:3890, text=mov_90_4)
df_mov_90_5 <- data_frame(line=1:7901, text=mov_90_5)
## 1980's:
df_mov_80_1 <- data_frame(line=1:3792, text=mov_80_1)
```

```
df_mov_80_2 <- data_frame(line=1:4017, text=mov_80_2)
df_mov_80_3 <- data_frame(line=1:3462, text=mov_80_3)
df_mov_80_4 <- data_frame(line=1:4735, text=mov_80_4)
df_mov_80_5 <- data_frame(line=1:3875, text=mov_80_5)
```

# Example



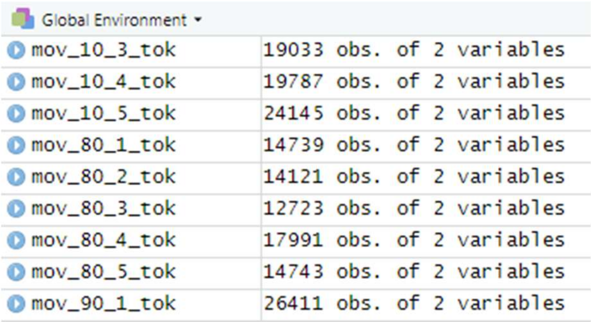
```
Data
df_mov_00_1    4993 obs. of 2 variables
df_mov_00_2    4628 obs. of 2 variables
df_mov_00_3    4822 obs. of 2 variables
df_mov_00_4    4033 obs. of 2 variables
df_mov_00_5    6253 obs. of 2 variables
df_mov_10_1    7945 obs. of 2 variables
df_mov_10_2    5896 obs. of 2 variables
df_mov_10_3    5307 obs. of 2 variables
```



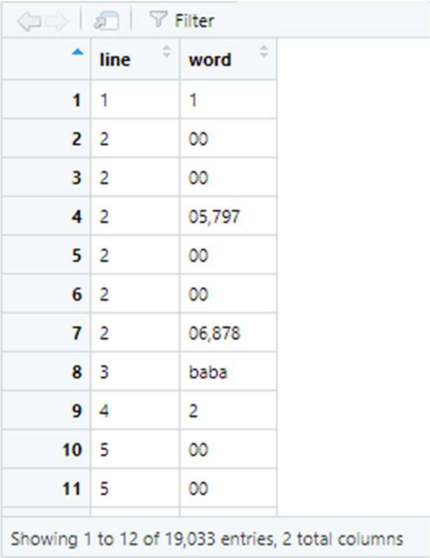
	line	text
1	1	ï»¿1
2	2	00:00:39,789 --> 00:00:42,124
3	3	When I was lying there in the VA hospital,
4	4	2
5	5	00:00:42,167 --> 00:00:45,127
6	6	with a big hole
7	7	blown through the middle of my life,
8	8	3
9	9	00:00:45,920 --> 00:00:48,422
10	10	I started having these dreams of flying.

Showing 1 to 12 of 4,993 entries, 2 total columns

```
# Tokenizing each dataframe: ####
mov_10_1_tok <- df_mov_10_1 %>%
  unnest_tokens(word,text)
mov_10_2_tok <- df_mov_10_2 %>%
  unnest_tokens(word,text)
mov_10_3_tok <- df_mov_10_3 %>%
  unnest_tokens(word,text)
mov_10_4_tok <- df_mov_10_4 %>%
  unnest_tokens(word,text)
mov_10_5_tok <- df_mov_10_5 %>%
  unnest_tokens(word,text)
mov_00_1_tok <- df_mov_00_1 %>%
  unnest_tokens(word,text)
mov_00_2_tok <- df_mov_00_2 %>%
  unnest_tokens(word,text)
mov_00_3_tok <- df_mov_00_3 %>%
  unnest_tokens(word,text)
mov_00_4_tok <- df_mov_00_4 %>%
  unnest_tokens(word,text)
mov_00_5_tok <- df_mov_00_5 %>%
  unnest_tokens(word,text)
mov_90_1_tok <- df_mov_90_1 %>%
  unnest_tokens(word,text)
mov_90_2_tok <- df_mov_90_2 %>%
  unnest_tokens(word,text)
mov_90_3_tok <- df_mov_90_3 %>%
  unnest_tokens(word,text)
mov_90_4_tok <- df_mov_90_4 %>%
  unnest_tokens(word,text)
mov_90_5_tok <- df_mov_90_5 %>%
  unnest_tokens(word,text)
mov_80_1_tok <- df_mov_80_1 %>%
  unnest_tokens(word,text)
mov_80_2_tok <- df_mov_80_2 %>%
  unnest_tokens(word,text)
mov_80_3_tok <- df_mov_80_3 %>%
  unnest_tokens(word,text)
mov_80_4_tok <- df_mov_80_4 %>%
  unnest_tokens(word,text)
mov_80_5_tok <- df_mov_80_5 %>%
  unnest_tokens(word,text)
# Example
```



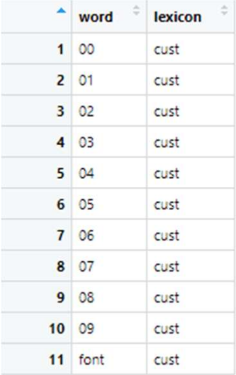
```
mov_10_3_tok    19033 obs. of 2 variables
mov_10_4_tok    19787 obs. of 2 variables
mov_10_5_tok    24145 obs. of 2 variables
mov_80_1_tok    14739 obs. of 2 variables
mov_80_2_tok    14121 obs. of 2 variables
mov_80_3_tok    12723 obs. of 2 variables
mov_80_4_tok    17991 obs. of 2 variables
mov_80_5_tok    14743 obs. of 2 variables
mov_90_1_tok    26411 obs. of 2 variables
```



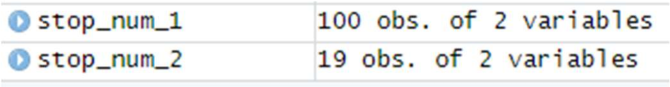
	line	word
1	1	1
2	2	00
3	2	00
4	2	05,797
5	2	00
6	2	00
7	2	06,878
8	3	baba
9	4	2
10	5	00
11	5	00

Showing 1 to 12 of 19,033 entries, 2 total columns

```
# Creating dictionary to remove numbers: ####
num_1 <- as.character(c(1:100))
stop_num_1 <- data_frame(word=num_1,lexicon="cust")
num_2 <- as.character(c("00","01","02","03","04","05","06","07","08","09",
                "font", "color", "ffff00", "ff0000",
                "ª", "â", "d900d9", "ff2424", "bb", "elliott", "e.t",
                "luke", "yoda", "batman", "alfred", "jack", "marcus",
                "leia", "jabba", "chewie", "r2", "rose", "simba",
                "pumba", "jedi", "david", "thanos", "rey", "hakuna",
                "mufasa", "zach", "hermione", "harry", "jake", "lord",
                "harvey", "tony", "gamora","groot", "stark", "claire",
                "gray","pumbaa","zazu","grace", "neytiri", "eywa","norm",
                "avatar", "potter","ron","hagrid","hogwarts","dumbledore",
                "sparrow","elizabeth","jones","turner","dent","wayne",
                "joker","frodo","agol","gondor","gandalf","mesa","scar",
                "dawson","gotham","mary","gertie","mike","vader","skywalker",
                "jones","solo", "hulk"))
stop_num_2 <- data_frame(word=num_2,lexicon="cust")
```



	word	lexicon
1	00	cust
2	01	cust
3	02	cust
4	03	cust
5	04	cust
6	05	cust
7	06	cust
8	07	cust
9	08	cust
10	09	cust
11	font	cust



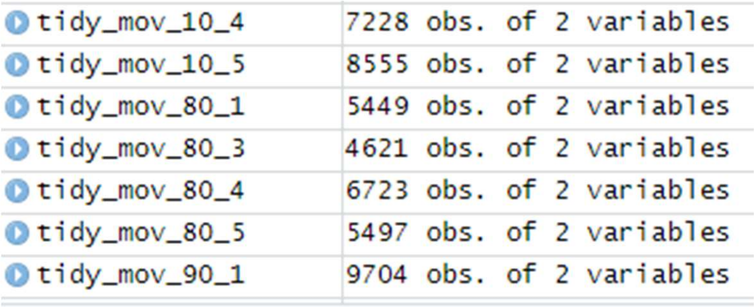
```
stop_num_1    100 obs. of 2 variables
stop_num_2    19 obs. of 2 variables
```

# Removing stop words per movie: ####

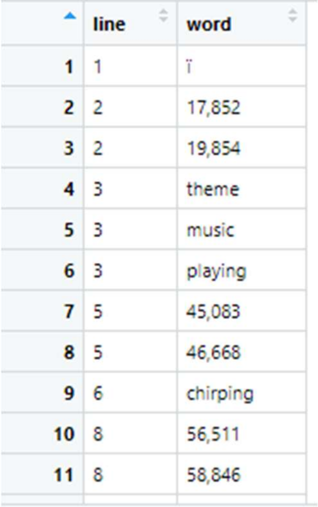
```
tidy_mov_10_1 <- df_mov_10_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_10_2 <- df_mov_10_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_10_3 <- df_mov_10_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_10_4 <- df_mov_10_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_10_5 <- df_mov_10_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_00_1 <- df_mov_00_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_00_2 <- df_mov_00_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_00_3 <- df_mov_00_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
```

```
tidy_mov_00_4 <- df_mov_00_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_00_5 <- df_mov_00_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_90_1 <- df_mov_90_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_90_2 <- df_mov_90_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_90_3 <- df_mov_90_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_90_4 <- df_mov_90_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_90_5 <- df_mov_90_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_80_1 <- df_mov_80_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_80_2 <- df_mov_80_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_80_3 <- df_mov_80_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_80_4 <- df_mov_80_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
tidy_mov_80_5 <- df_mov_80_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2)
```

# Example



```
tidy_mov_10_4    7228 obs. of 2 variables
tidy_mov_10_5    8555 obs. of 2 variables
tidy_mov_80_1    5449 obs. of 2 variables
tidy_mov_80_3    4621 obs. of 2 variables
tidy_mov_80_4    6723 obs. of 2 variables
tidy_mov_80_5    5497 obs. of 2 variables
tidy_mov_90_1    9704 obs. of 2 variables
```

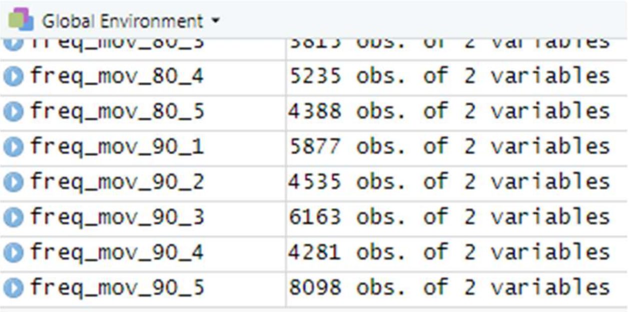


	line	word
1	1	ï
2	2	17,852
3	2	19,854
4	3	theme
5	3	music
6	3	playing
7	5	45,083
8	5	46,668
9	6	chirping
10	8	56,511
11	8	58,846

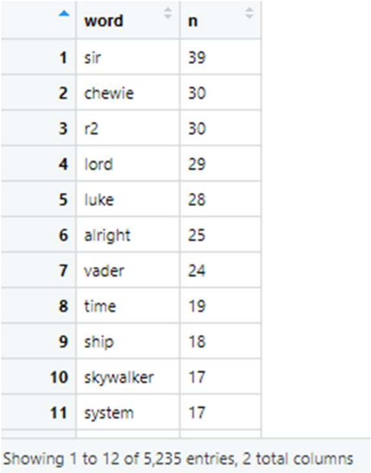
# Counting frequency per movie: ####

```
freq_mov_10_1 <- df_mov_10_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_10_2 <- df_mov_10_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_10_3 <- df_mov_10_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_10_4 <- df_mov_10_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_10_5 <- df_mov_10_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_00_1 <- df_mov_00_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_00_2 <- df_mov_00_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_00_3 <- df_mov_00_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_00_4 <- df_mov_00_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_00_5 <- df_mov_00_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_90_1 <- df_mov_90_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_90_2 <- df_mov_90_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_90_3 <- df_mov_90_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_90_4 <- df_mov_90_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_90_5 <- df_mov_90_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_80_1 <- df_mov_80_1 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_80_2 <- df_mov_80_2 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_80_3 <- df_mov_80_3 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_80_4 <- df_mov_80_4 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
freq_mov_80_5 <- df_mov_80_5 %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  anti_join(stop_num_1) %>%
  anti_join(stop_num_2) %>%
  count(word, sort=TRUE)
```

# Example



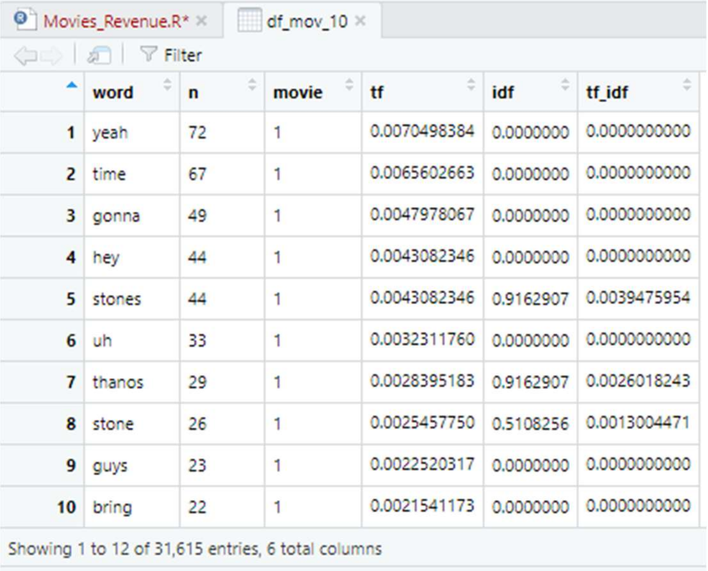
```
freq_mov_80_3    5815 obs. of 2 variables
freq_mov_80_4    5235 obs. of 2 variables
freq_mov_80_5    4388 obs. of 2 variables
freq_mov_90_1    5877 obs. of 2 variables
freq_mov_90_2    4535 obs. of 2 variables
freq_mov_90_3    6163 obs. of 2 variables
freq_mov_90_4    4281 obs. of 2 variables
freq_mov_90_5    8098 obs. of 2 variables
```



	word	n
1	sir	39
2	chewie	30
3	r2	30
4	lord	29
5	luke	28
6	alright	25
7	vader	24
8	time	19
9	ship	18
10	skywalker	17
11	system	17

Showing 1 to 12 of 5,235 entries, 2 total columns

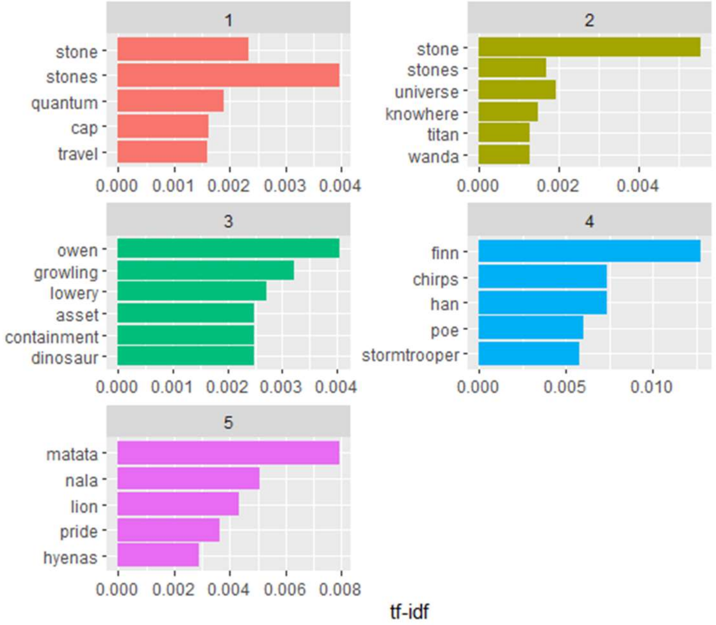
```
# Sentiments: ####
afinn <- get_sentiments("afinn")
nrc <- get_sentiments("nrc")
bing <- get_sentiments("bing")

sentiments <- bind_rows(mutate(afinn, lexicon="afinn"),
                        mutate(nrc, lexicon= "nrc"),
                        mutate(bing, lexicon="bing")
)
# Analyzing frequency per decade (binding): ####
# Creating empty dataframe:
a <- 7945
b <- 3
df_mov_10 <- as.data.frame(matrix(nrow=a, ncol=b))
c <- 6253
df_mov_00 <- as.data.frame(matrix(nrow=c, ncol=b))
d <- 7901
df_mov_90 <- as.data.frame(matrix(nrow=d, ncol=b))
e <- 4735
df_mov_80 <- as.data.frame(matrix(nrow=e, ncol=b))
# Binding rows:
df_mov_10 <- bind_rows(
  mutate(freq_mov_10_1, movie='1'),
  mutate(freq_mov_10_2, movie='2'),
  mutate(freq_mov_10_3, movie='3'),
  mutate(freq_mov_10_4, movie='4'),
  mutate(freq_mov_10_5, movie='5')
)
df_mov_00 <- bind_rows(
  mutate(freq_mov_00_1, movie='1'),
  mutate(freq_mov_00_2, movie='2'),
  mutate(freq_mov_00_3, movie='3'),
  mutate(freq_mov_00_4, movie='4'),
  mutate(freq_mov_00_5, movie='5')
)
df_mov_90 <- bind_rows(
  mutate(freq_mov_90_1, movie='1'),
  mutate(freq_mov_90_2, movie='2'),
  mutate(freq_mov_90_3, movie='3'),
  mutate(freq_mov_90_4, movie='4'),
  mutate(freq_mov_90_5, movie='5')
)
df_mov_80 <- bind_rows(
  mutate(freq_mov_80_1, movie='1'),
  mutate(freq_mov_80_2, movie='2'),
  mutate(freq_mov_80_3, movie='3'),
  mutate(freq_mov_80_4, movie='4'),
  mutate(freq_mov_80_5, movie='5')
)
# Including proportion per word:
df_mov_10 <- df_mov_10 %>%
  bind_tf_idf(word, movie, n)
df_mov_00 <- df_mov_00 %>%
  bind_tf_idf(word, movie, n)
df_mov_90 <- df_mov_90 %>%
  bind_tf_idf(word, movie, n)
df_mov_80 <- df_mov_80 %>%
  bind_tf_idf(word, movie, n)
# Example
```



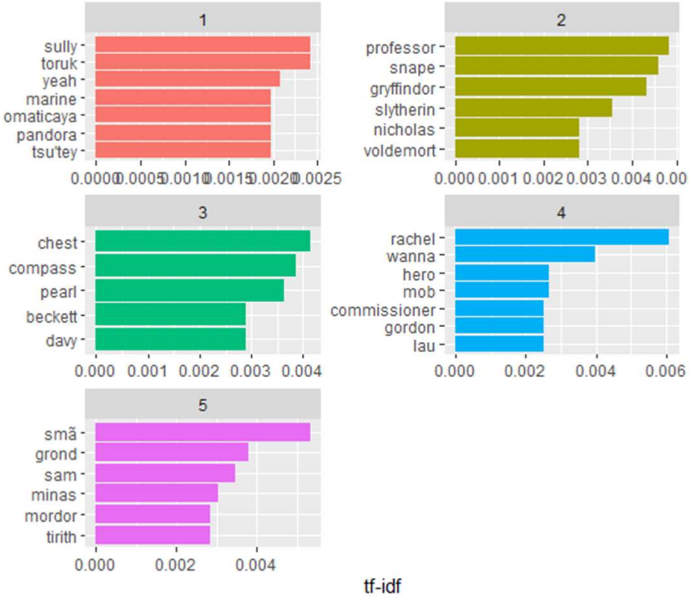
	word	n	movie	tf	idf	tf_idf
1	yeah	72	1	0.0070498384	0.0000000	0.0000000000
2	time	67	1	0.0065602663	0.0000000	0.0000000000
3	gonna	49	1	0.0047978067	0.0000000	0.0000000000
4	hey	44	1	0.0043082346	0.0000000	0.0000000000
5	stones	44	1	0.0043082346	0.9162907	0.0039475954
6	uh	33	1	0.0032311760	0.0000000	0.0000000000
7	thanos	29	1	0.0028395183	0.9162907	0.0026018243
8	stone	26	1	0.0025457750	0.5108256	0.0013004471
9	guys	23	1	0.0022520317	0.0000000	0.0000000000
10	bring	22	1	0.0021541173	0.0000000	0.0000000000

Showing 1 to 12 of 31,615 entries, 6 total columns

```
# Plotting frequecy (per movie per decade): ####
# 2010
df_mov_10 %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(movie) %>%
  top_n(5) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=movie))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~movie, ncol=2, scales="free")+
  coord_flip()
```

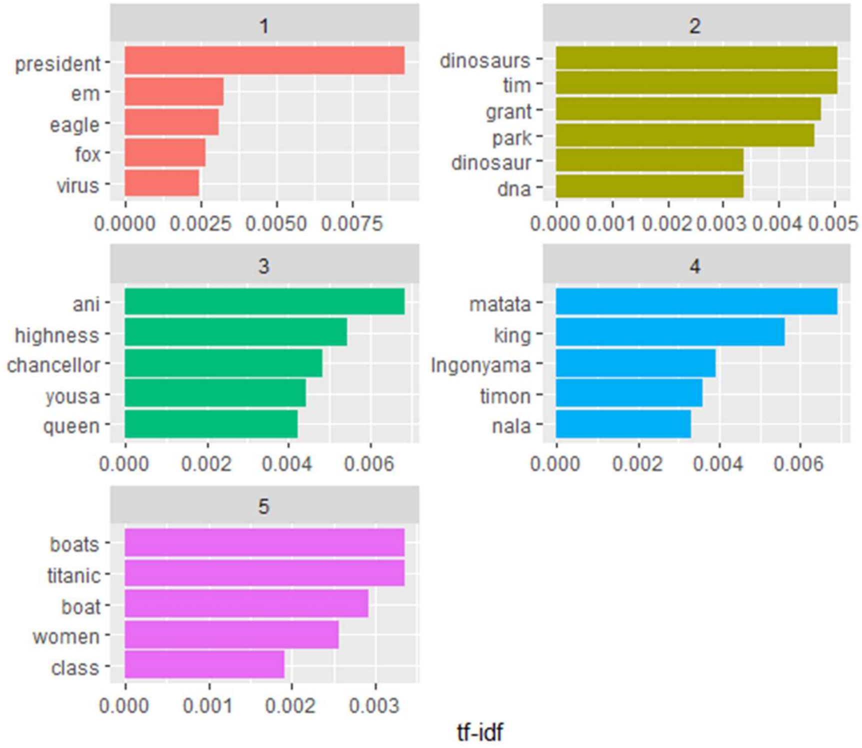


```
# 2000
df_mov_00 %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(movie) %>%
  top_n(5) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=movie))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~movie, ncol=2, scales="free")+
  coord_flip()
```

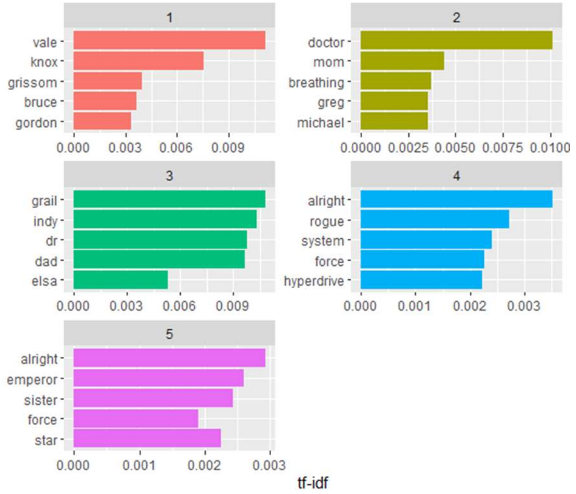


# 1990

```
df_mov_90 %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(movie) %>%
  top_n(5) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=movie))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~movie, ncol=2, scales="free")+
  coord_flip()
```



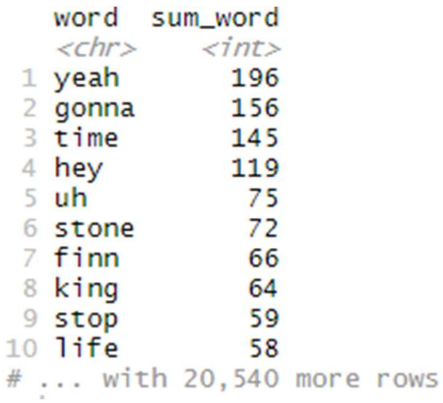
```
# 1980
df_mov_80 %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(movie) %>%
  top_n(5) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=movie))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~movie, ncol=2, scales="free")+
  coord_flip()
```



# Consolidating analysis per decade: ####

## Frequency consolidating top 5 2010's movies:

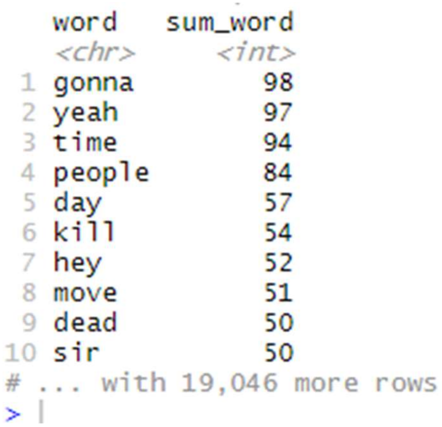
```
cons_freq_10 <- df_mov_10 %>%
  group_by(word) %>%
  summarise(sum_word = sum(n)) %>%
  arrange(desc(sum_word))
cons_freq_10
```



```
   word   sum_word
   <chr>     <int>
 1 yeah        196
 2 gonna       156
 3 time        145
 4 hey         119
 5 uh           75
 6 stone        72
 7 finn         66
 8 king         64
 9 stop         59
10 life         58
# ... with 20,540 more rows
```

## Frequency consolidating top 5 2000's movies:

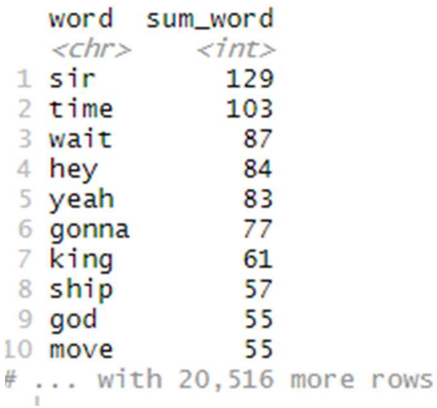
```
cons_freq_00 <- df_mov_00 %>%
  group_by(word) %>%
  summarise(sum_word = sum(n)) %>%
  arrange(desc(sum_word))
cons_freq_00
```



```
   word    sum_word
   <chr>      <int>
 1 gonna         98
 2 yeah          97
 3 time          94
 4 people        84
 5 day           57
 6 kill          54
 7 hey           52
 8 move          51
 9 dead          50
10 sir           50
# ... with 19,046 more rows
> |
```

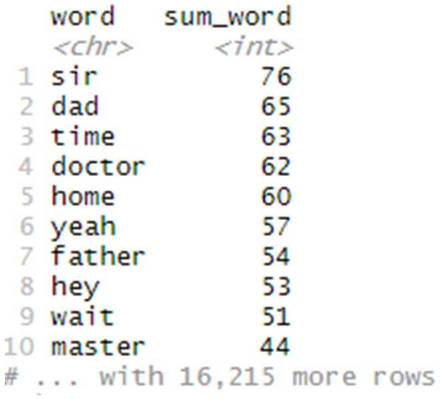
## Frequency consolidating top 5 1990's movies:

```
cons_freq_90 <- df_mov_90 %>%
  group_by(word) %>%
  summarise(sum_word = sum(n)) %>%
  arrange(desc(sum_word))
cons_freq_90
```



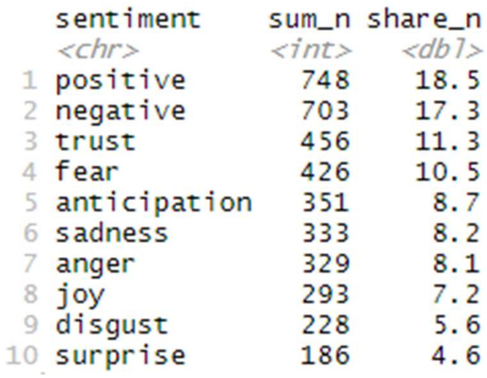
```
   word   sum_word
   <chr>     <int>
 1 sir         129
 2 time        103
 3 wait         87
 4 hey          84
 5 yeah         83
 6 gonna        77
 7 king         61
 8 ship         57
 9 god          55
10 move         55
# ... with 20,516 more rows
```

```
## Frequency consolidating top 5 1980's movies:
cons_freq_80 <- df_mov_80 %>%
  group_by(word) %>%
  summarise(sum_word = sum(n)) %>%
  arrange(desc(sum_word))
cons_freq_80
```



```
   word    sum_word
   <chr>      <int>
 1 sir           76
 2 dad           65
 3 time          63
 4 doctor        62
 5 home          60
 6 yeah          57
 7 father        54
 8 hey           53
 9 wait          51
10 master        44
# ... with 16,215 more rows
```

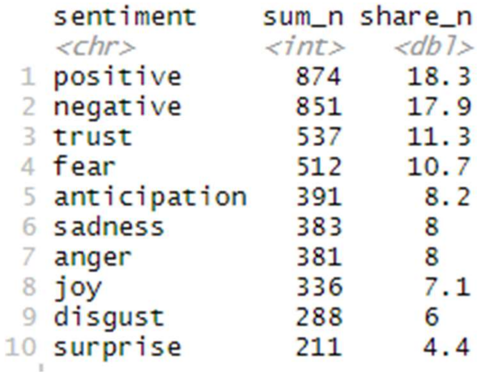
```
# Getting frequency of sentiments per decade: ####
## NRC | 2010's movies
nrc_mov_10 <- df_mov_10 %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
sent_freq_10<-nrc_mov_10 %>%
  group_by(sentiment) %>%
  summarise(sum_n = sum(n),
            share_n = round((sum_n/4053)*100,digits=1))%>%
  arrange(desc(sum_n))
sent_freq_10
```



```
   sentiment    sum_n share_n
   <chr>        <int>   <dbl>
 1 positive       748    18.5
 2 negative       703    17.3
 3 trust          456    11.3
 4 fear           426    10.5
 5 anticipation   351     8.7
 6 sadness        333     8.2
 7 anger          329     8.1
 8 joy            293     7.2
 9 disgust        228     5.6
10 surprise       186     4.6
```

## NRC | 2000's movies

```
nrc_mov_00 <- df_mov_00 %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
sent_freq_00<-nrc_mov_00 %>%
  group_by(sentiment) %>%
  summarise(sum_n = sum(n),
            share_n = round((sum_n/4764)*100,digits=1))%>%
  arrange(desc(sum_n))
sent_freq_00
```

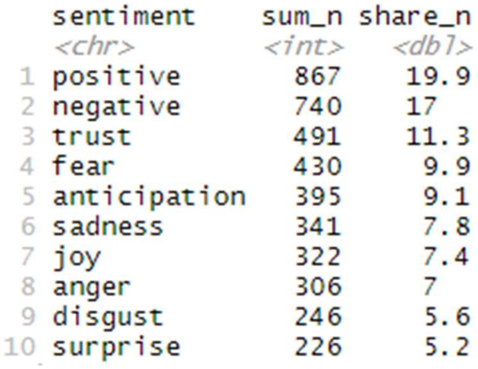


```
   sentiment    sum_n share_n
   <chr>        <int>   <dbl>
 1 positive       874    18.3
 2 negative       851    17.9
 3 trust          537    11.3
 4 fear           512    10.7
 5 anticipation   391     8.2
 6 sadness        383     8
 7 anger          381     8
 8 joy            336     7.1
 9 disgust        288     6
10 surprise       211     4.4
```

## NRC | 1990's movies

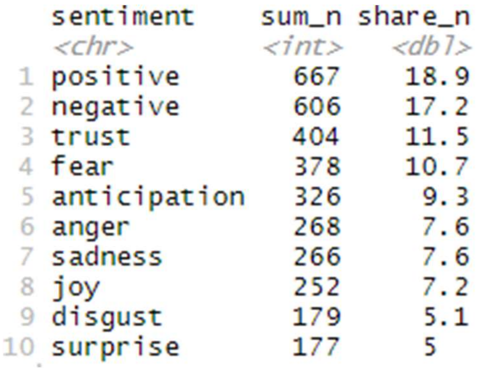
```
nrc_mov_90 <- df_mov_90 %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

sent_freq_90<-nrc_mov_90 %>%
  group_by(sentiment) %>%
  summarise(sum_n = sum(n),
            share_n = round((sum_n/4364)*100,digits=1))%>%
  arrange(desc(sum_n))
```

```
sent_freq_90
```



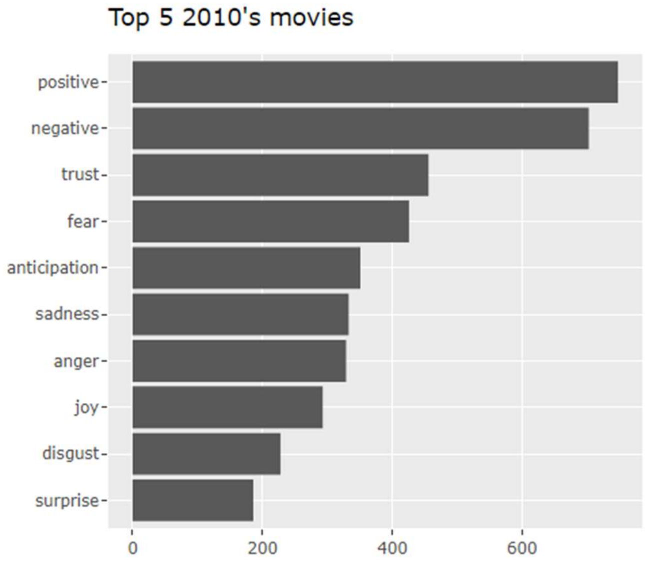
```
   sentiment    sum_n share_n
   <chr>        <int>   <dbl>
 1 positive       867    19.9
 2 negative       740    17  
 3 trust          491    11.3
 4 fear           430     9.9
 5 anticipation   395     9.1
 6 sadness        341     7.8
 7 joy            322     7.4
 8 anger          306     7  
 9 disgust        246     5.6
10 surprise       226     5.2
```

```
## NRC | 1980's movies
nrc_mov_80 <- df_mov_80 %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

sent_freq_80<-nrc_mov_80 %>%
  group_by(sentiment) %>%
  summarise(sum_n = sum(n),
            share_n = round((sum_n/3523)*100,digits=1))%>%
  arrange(desc(sum_n))
sent_freq_80
```



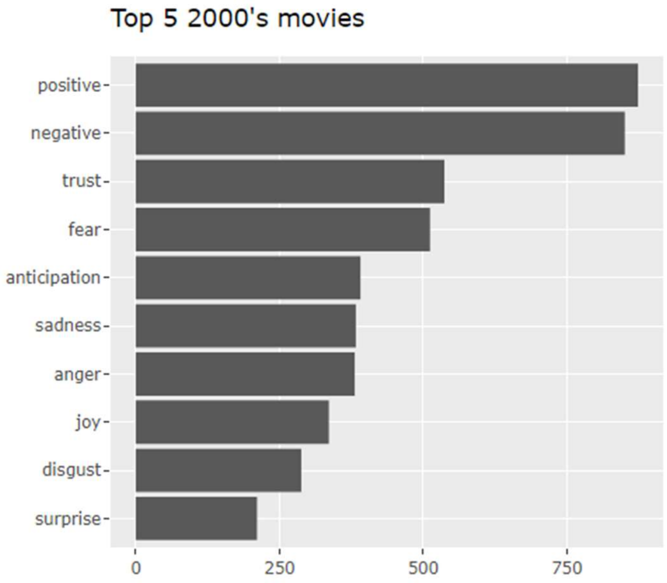
```
   sentiment    sum_n share_n
   <chr>        <int>   <dbl>
 1 positive       667    18.9
 2 negative       606    17.2
 3 trust          404    11.5
 4 fear           378    10.7
 5 anticipation   326     9.3
 6 anger          268     7.6
 7 sadness        266     7.6
 8 joy            252     7.2
 9 disgust        179     5.1
10 surprise       177     5  
```

```
# Plotting sentiments frequency per decade: ####
# Barplot (ranking of NRC sentiments): ####
## 2010's movies
plot_sent_10 <-sent_freq_10 %>%
  mutate(sentiment = reorder(sentiment,sum_n)) %>%
  ggplot(aes(sentiment, sum_n))+
  geom_col()+
  xlab(NULL)+
  ylab(NULL)+
  coord_flip() +
  ggtitle("Top 5 2010's movies")
ggplotly(plot_sent_10)
```

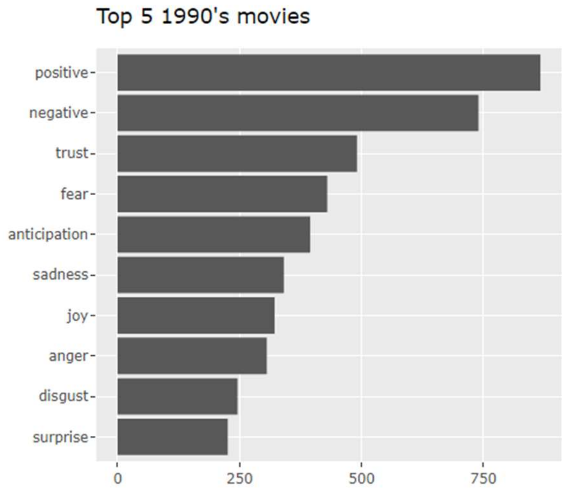


## 2000's movies

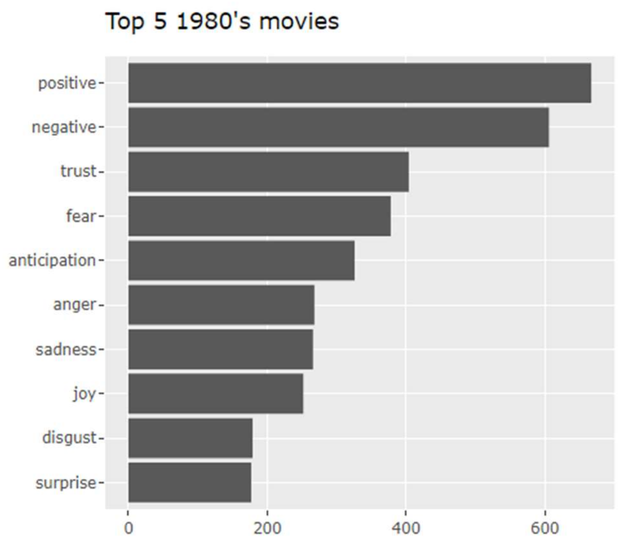
```
plot_sent_00 <-sent_freq_00 %>%
  mutate(sentiment = reorder(sentiment,sum_n)) %>%
  ggplot(aes(sentiment, sum_n))+
  geom_col()+
  xlab(NULL)+
  ylab(NULL)+
  coord_flip() +
  ggtitle("Top 5 2000's movies")
ggplotly(plot_sent_00)
```



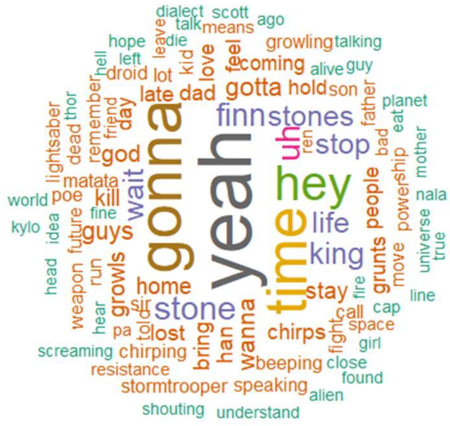
```
## 1990's movies
plot_sent_90 <-sent_freq_90 %>%
  mutate(sentiment = reorder(sentiment,sum_n)) %>%
  ggplot(aes(sentiment, sum_n))+
  geom_col()+
  xlab(NULL)+
  ylab(NULL)+
  coord_flip() +
  ggtitle("Top 5 1990's movies")
ggplotly(plot_sent_90)
```



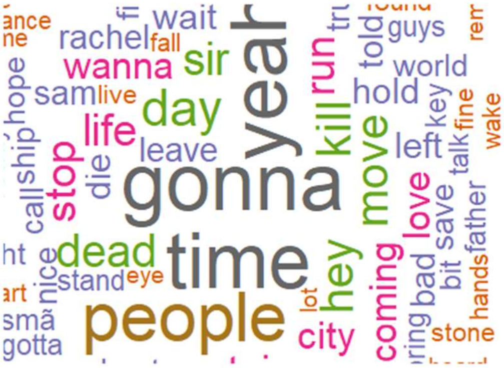
```
## 1980's movies
plot_sent_80 <-sent_freq_80 %>%
  mutate(sentiment = reorder(sentiment,sum_n)) %>%
  ggplot(aes(sentiment, sum_n))+
  geom_col()+
  xlab(NULL)+
  ylab(NULL)+
  coord_flip() +
  ggtitle("Top 5 1980's movies")
ggplotly(plot_sent_80)
```



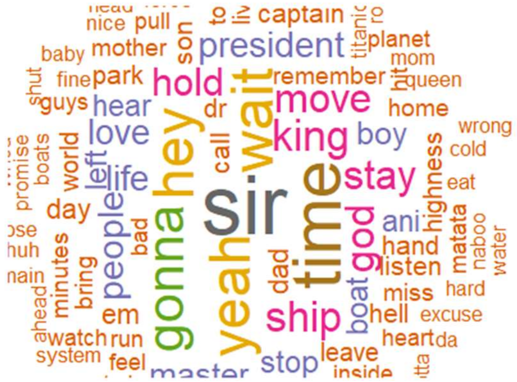
```
# Wordcloud (consolidated per decade): ####
# 2010
cons_freq_10 %>%
  with(wordcloud(word, sum_word, max.words = 100, scale=c(5,0.5),
               random.order=FALSE, rot.per=0.35, use.r.layout=FALSE,
               colors=brewer.pal(8,"Dark2")))
```



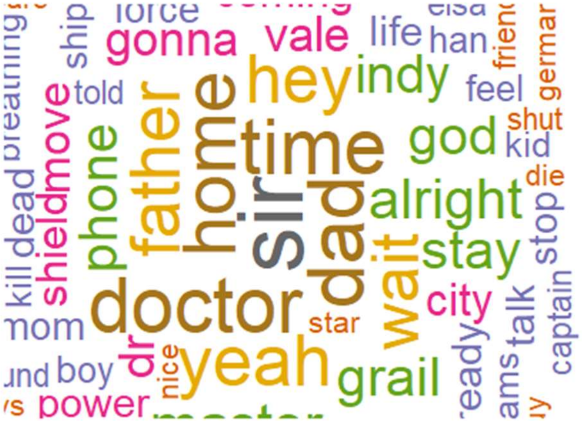
```
# 2000
cons_freq_00 %>%
  with(wordcloud(word, sum_word, max.words = 100, scale=c(5,0.5),
              random.order=FALSE, rot.per=0.35, use.r.layout=FALSE,
              colors=brewer.pal(8,"Dark2")))
```



```
# 1990
cons_freq_90 %>%
  with(wordcloud(word, sum_word, max.words = 100, scale=c(5,0.5),
              random.order=FALSE, rot.per=0.35, use.r.layout=FALSE,
              colors=brewer.pal(8,"Dark2")))
```



```
# 1980
cons_freq_80 %>%
  with(wordcloud(word, sum_word, max.words = 100, scale=c(5,0.5),
              random.order=FALSE, rot.per=0.35, use.r.layout=FALSE,
              colors=brewer.pal(8,"Dark2")))
```


sir
time
doctor
dad
home
father
yeah
hey
wait
indy
alright
god
stay
grail
phone
gonna
vale
city
move
shield
dead
kill
mom
boy
power
dr
nice
star
told
ship
force
life
han
feel
shut
kid
die
stop
talk
captain
elsa