Harry Potter 1-3 EDA

```
options(warn = -23) # ignore all warnings
options(scipen = 10000)
options(repr.plot.width = 14.0, repr.plot.height = 10.0)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
#library(magrittr)
library(scales) # visualisation
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(RColorBrewer) # color visualisation
library(ggsci)
library(ggthemes)
library(lubridate) # date ant time management
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(viridis) # color maps
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:scales':
```

```
##
##     viridis_pal
```

```
library(ggrepel)
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:lubridate':
##
##     stamp

## The following object is masked from 'package:dplyr':
##
##     rename

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

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(tm) # text mining
```

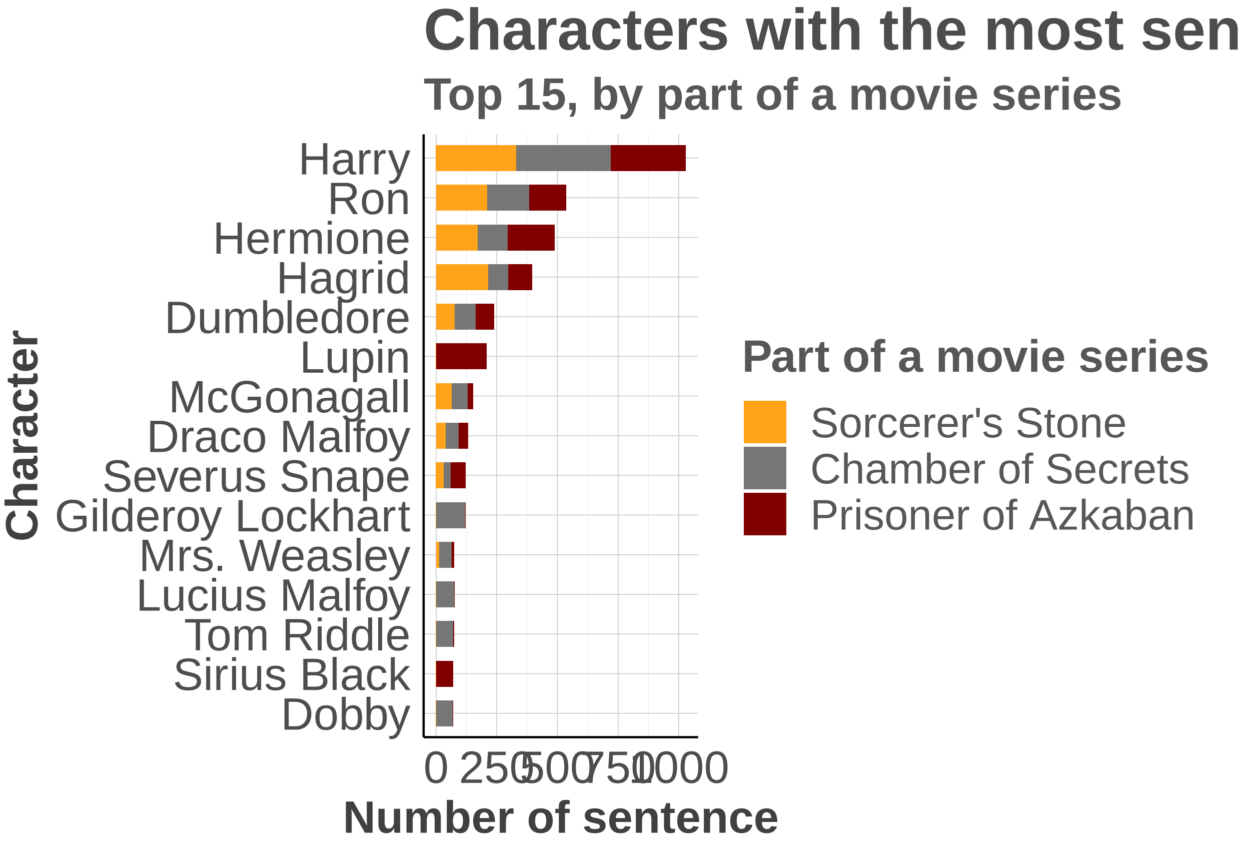
```
## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(SnowballC) #snowball stemmer
library(wordcloud)
library(NLP)
library(widyr)
library(wordcloud2)
library(tidytext)
library(janeaustenr)
library(htmlwidgets)

annotate <- ggplot2::annotate

theme_michau <- theme(legend.position = "bottom", legend.direction = "horizontal", axis.text = element_
plot.caption = element_text(color = "gray65", size = 12.4), legend.text = element_text(size = 16, colou
axis.title = element_text(size = 16.7, face = "bold", color = "gray25"), legend.title = element_text(si
axis.line = element_line(size = 0.4), plot.title = element_text(size = 21.9, face = "bold", colour = "g
panel.grid.major = element_line(colour = "gray80", size = 0.15), plot.subtitle = element_text(size = 16
strip.text = element_text(size = 16.7, face = "bold"), panel.grid.minor = element_line(size = 0))
```

Input all the data from films 1-3.

```
Script1 <- read.csv("harry-potter-dataset/Harry Potter 1.csv", row.names = NULL, sep = ";", encoding =
Script2 <- read.csv("harry-potter-dataset/Harry Potter 2.csv", row.names = NULL, sep = ";", encoding =
Script3 <- read.csv("harry-potter-dataset/Harry Potter 3.csv", row.names = NULL, sep = ";", encoding =

names(Script3) <- c("Character", "Sentence")

Script1$Character <- as.character(str_trim(Script1$Character, side = "both"))
Script2$Character <- as.character(str_trim(Script2$Character, side = "both"))
Script3$Character <- as.character(str_trim(Script3$Character, side = "both"))

Script1$Part <- "Sorcerer's Stone"
Script2$Part <- "Chamber of Secrets"
Script3$Part <- "Prisoner of Azkaban"

Script <- rbind(Script1, Script2, Script3)

Script$Part <- factor(Script$Part, levels=c("Prisoner of Azkaban", "Chamber of Secrets", "Sorcerer's Sto
Script$Character <- str_to_title(Script$Character)

Script <- Script %>%
  mutate(Character = case_when(Character %in% c("Dumbledore") ~ "Dumbledore",
                               Character %in% c("Mcgonagall") ~ "McGonagall",
                               Character %in% c("Hagrid") ~ "Hagrid",
                               Character %in% c("Petunia", "Aunt Petunia") ~ "Aunt Petunia",
                               Character %in% c("Dudley") ~ "Dudley",
                               Character %in% c("Vernon") ~ "Vernon",
                               Character %in% c("Harry") ~ "Harry",
                               Character %in% c("Snake") ~ "Snake",
                               Character %in% c("Someone") ~ "Someone",
                               Character %in% c("Barkeep Tom") ~ "Barkeep Tom",
                               Character %in% c("Man", "Boy", "Boy 1", "Boy 2") ~ "Man/Boy",
                               Character %in% c("Witch") ~ "Witch",
                               Character %in% c("Quirrell") ~ "Quirrell",
                               Character %in% c("Goblin") ~ "Goblin",
                               Character %in% c("Griphook") ~ "Griphook",
                               Character %in% c("Ollivander") ~ "Ollivander",
                               Character %in% c("Trainmaster") ~ "Trainmaster",
                               Character %in% c("Mrs. Weasley") ~ "Mrs. Weasley",
                               Character %in% c("George") ~ "George",
                               Character %in% c("Fred") ~ "Fred",
                               Character %in% c("Ginny") ~ "Ginny",
                               Character %in% c("Ron") ~ "Ron",
                               Character %in% c("Woman", "Girl") ~ "Girl/Woman",
                               Character %in% c("Hermione", "Hermoine") ~ "Hermione",
                               Character %in% c("Neville") ~ "Neville",
                               Character %in% c("Malfoy", "Draco") ~ "Draco Malfoy",
                               Character %in% c("Sorting Hat") ~ "Sorting Hat",
                               Character %in% c("Seamus") ~ "Seamus",
                               Character %in% c("Percy") ~ "Percy",
                               Character %in% c("Sir Nicholas") ~ "Sir Nicholas",
                               Character %in% c("Man In Paint") ~ "Man In Paint",
                               Character %in% c("Fat Lady") ~ "Fat Lady",
                               Character %in% c("Snape") ~ "Severus Snape",
                               Character %in% c("Dean") ~ "Dean",
                               Character %in% c("Madam Hooch") ~ "Madam Hooch",
                               Character %in% c("Filch") ~ "Filch",
                               Character %in% c("All", "All 3") ~ "Crowd",
                               Character %in% c("Lee Jordan", "Lee  Jordan") ~ "Lee Jordan",
                               Character %in% c("Gryffindors") ~ "Gryffindors",
                               Character %in% c("Flint") ~ "Flint",
                               Character %in% c("Firenze") ~ "Firenze",
                               Character %in% c("Voldemort") ~ "Voldemort",
                               Character %in% c("Students", "Student", "Class") ~ "Student",
                               Character %in% c("Crowd") ~ "Crowd",
                               Character %in% c("Uncle Vernon") ~ "Uncle Vernon",
                               Character %in% c("Dobby") ~ "Dobby",
                               Character %in% c("Aunt Petunia & Dudley") ~ "Aunt Petunia & Dudley",
                               Character %in% c("Mr. Weasley") ~ "Mr. Weasley",
                               Character %in% c("Fred, George, Ron") ~ "Fred, George, Ron",
                               Character %in% c("Fred, George, Ron, Harry") ~ "Fred, George, Ron, Harry"
                               Character %in% c("Lucius Malfoy") ~ "Lucius Malfoy",
                               Character %in% c("Photographer") ~ "Photographer",
                               Character %in% c("Lockhart", "Gilderoy Lockhart") ~ "Gilderoy Lockhart",
                               Character %in% c("Harry And Ron") ~ "Harry And Ron",
                               Character %in% c("Professor Sprout") ~ "Professor Sprout",
                               Character %in% c("Penelope Clearwater") ~ "Penelope Clearwater",
                               Character %in% c("Colin") ~ "Colin",
                               Character %in% c("Cornish Pixies") ~ "Cornish Pixies",
                               Character %in% c("Wood", "Oliver") ~ "Oliver Wood",
                               Character %in% c("Voice") ~ "Voice",
                               Character %in% c("Lupin") ~ "Lupin",
                               Character %in% c("Picture") ~ "Picture",
                               Character %in% c("Slytherins") ~ "Slytherins",
                               Character %in% c("Madam Pomfrey") ~ "Madam Pomfrey",
                               Character %in% c("Moaning Myrtle") ~ "Moaning Myrtle",
                               Character %in% c("Justin Finch-Fletchley") ~ "Justin Finch-Fletchley",
                               Character %in% c("Crabbe") ~ "Crabbe",
                               Character %in% c("Diary") ~ "Diary",
                               Character %in% c("Tom Riddle", "Tom") ~ "Tom Riddle",
                               Character %in% c("Harry-Ron-Hermione") ~ "Harry & Ron and Hermione",
                               Character %in% c("Fudge") ~ "Cornelius Fudge",
                               Character %in% c("Aragog") ~ "Aragog",
                               Character %in% c("Aunt Marge") ~ "Aunt Marge",
                               Character %in% c("Stan Shunpike") ~ "Stan Shunpike",
                               Character %in% c("Vendor") ~ "Vendor",
                               Character %in% c("Housekeeper") ~ "Housekeeper",
                               Character %in% c("Trelawney") ~ "Sybilla Trelawney",
                               Character %in% c("Bem") ~ "Bem",
                               Character %in% c("Pansy Parkinson") ~ "Pansy Parkinson",
                               Character %in% c("Parvati") ~ "Parvati Patil",
                               Character %in% c("Teacher") ~ "Teacher",
                               Character %in% c("Fred & George") ~ "Fred and George",
                               Character %in% c("Madam Rosmerta") ~ "Madam Rosmerta",
                               Character %in% c("Shrunken Head", "Shrunken Head 1", "Shrunken Head 2") ~
                               Character %in% c("Goyle") ~ "Goyle",
                               Character %in% c("Sirius") ~ "Sirius Black",
                               Character %in% c("Pettigrew") ~ "Peter Pettigrew"))


 Bing <- get_sentiments("bing")

 firstup <- function(x) {
   substr(x, 1, 1) <- toupper(substr(x, 1, 1))
   x
 }

 Bing$sentiment <- firstup(Bing$sentiment)
```

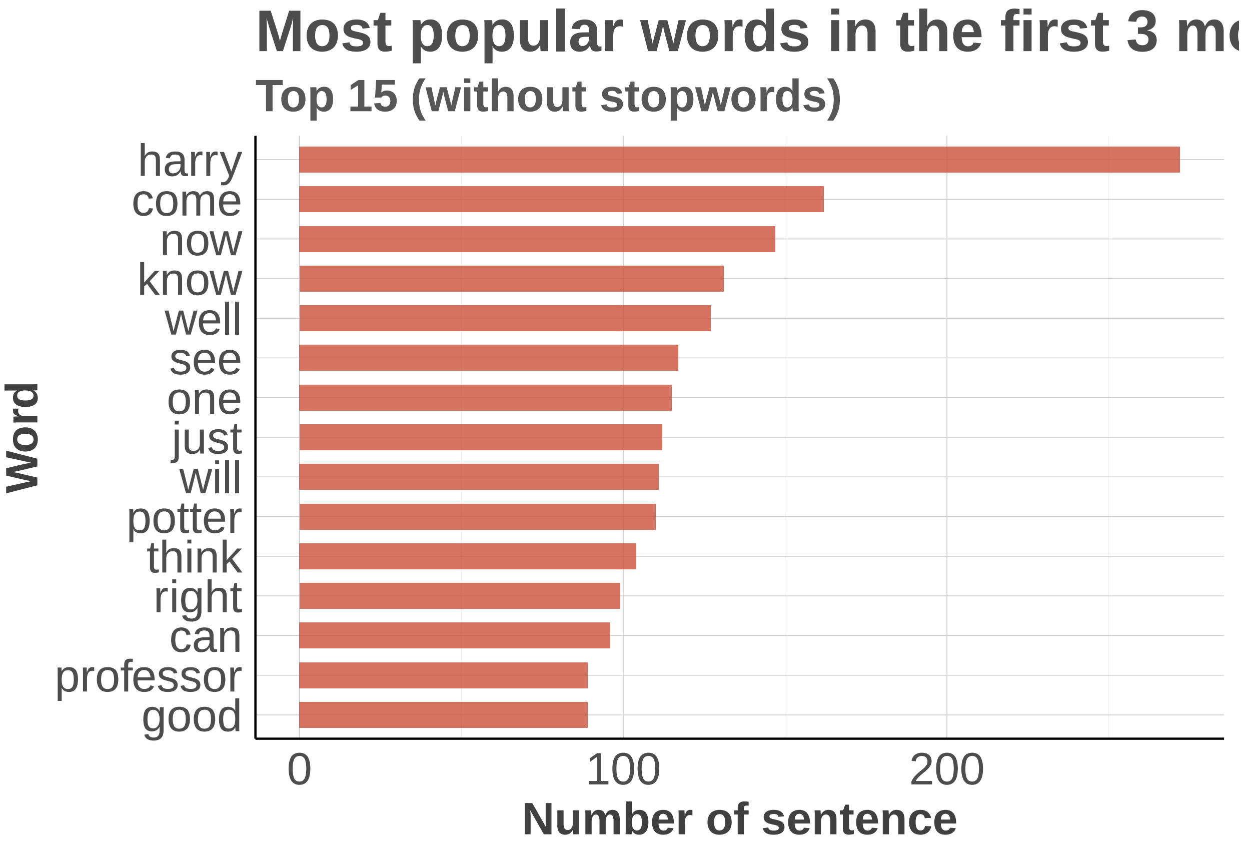
```
Char_Dial <- data.frame(table(Script$Character, Script$Part))
Char_Dial %>%
  arrange(desc(Freq)) %>%
  filter(Var1 %in% c("Harry", "Ron", "Hermione", "Hagrid", "Dumbledore", "Lupin", "McGonagall", "Draco 
                     "Severus Snape", "Lucius Malfoy", "Mrs. Weasley", "Tom Riddle", "Sirius Black", "De
  ggplot(., aes(reorder(Var1, +Freq), Freq, fill = Var2))+
  geom_bar(stat = "identity", width = 0.65)+
  scale_fill_uchicago()+
  coord_flip()+
  guides(fill = guide_legend(title.position = "top", reverse = T))+
  labs(title = "Characters with the most sentences",
       subtitle = "Top 15, by part of a movie series", fill = "Part of a movie series",
       x = "Character", y = "Number of sentence")+
  theme_minimal()+
  theme_michau+
  theme(legend.title.align = 0.5, legend.position = "right", legend.direction = "vertical")
```



```r
tm <- Corpus(VectorSource(Script$Sentence))
tm <- tm_map(tm, content_transformer(tolower))
tm <- tm_map(tm, removeNumbers)
tm <- tm_map(tm, removeWords, stopwords("english"))
tm <- tm_map(tm, removePunctuation)
tm <- tm_map(tm, stripWhitespace)
tdm <- TermDocumentMatrix(tm)

tdm <- as.matrix(tdm)
tdm <- sort(rowSums(tdm), decreasing = T)
tdm <- data.frame(Word = names(tdm), Number = tdm)
```

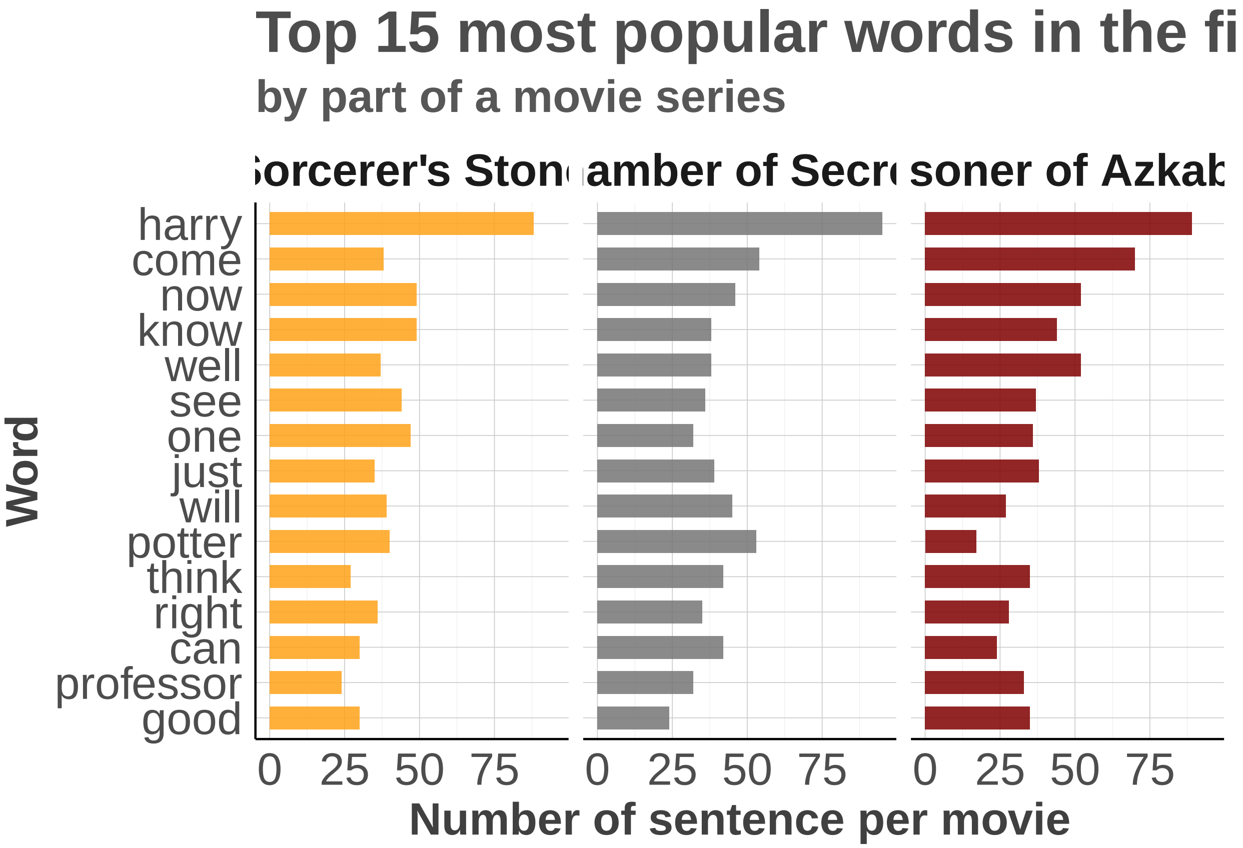
```r
wc <- tdm %>%
  filter(Number > 8) %>%
  select(Word, Number) %>%
  wordcloud2(., color = alpha("coral3", seq(0.9,0.2,-0.002)), backgroundColor = "white", size = 0.9)

saveWidget(wc,'wordcloud2.html',selfcontained = F)
set.seed(111)
IRdisplay::display_html('<iframe src="wordcloud2.html" width=99% height=500></iframe>')
```

```r
tdm %>%
  arrange(desc(Number)) %>%
  slice(1:15) %>%
ggplot(., aes(reorder(Word, +Number), Number))+
  geom_bar(stat = "identity", width = 0.65, fill = "coral3", alpha = 0.85)+
  coord_flip()+
  labs(title = "Most popular words in the first 3 movies",
      subtitle = "Top 15 (without stopwords)",
      x = "Word", y = "Number of sentence")+
  theme_minimal()+
  theme_michau
```

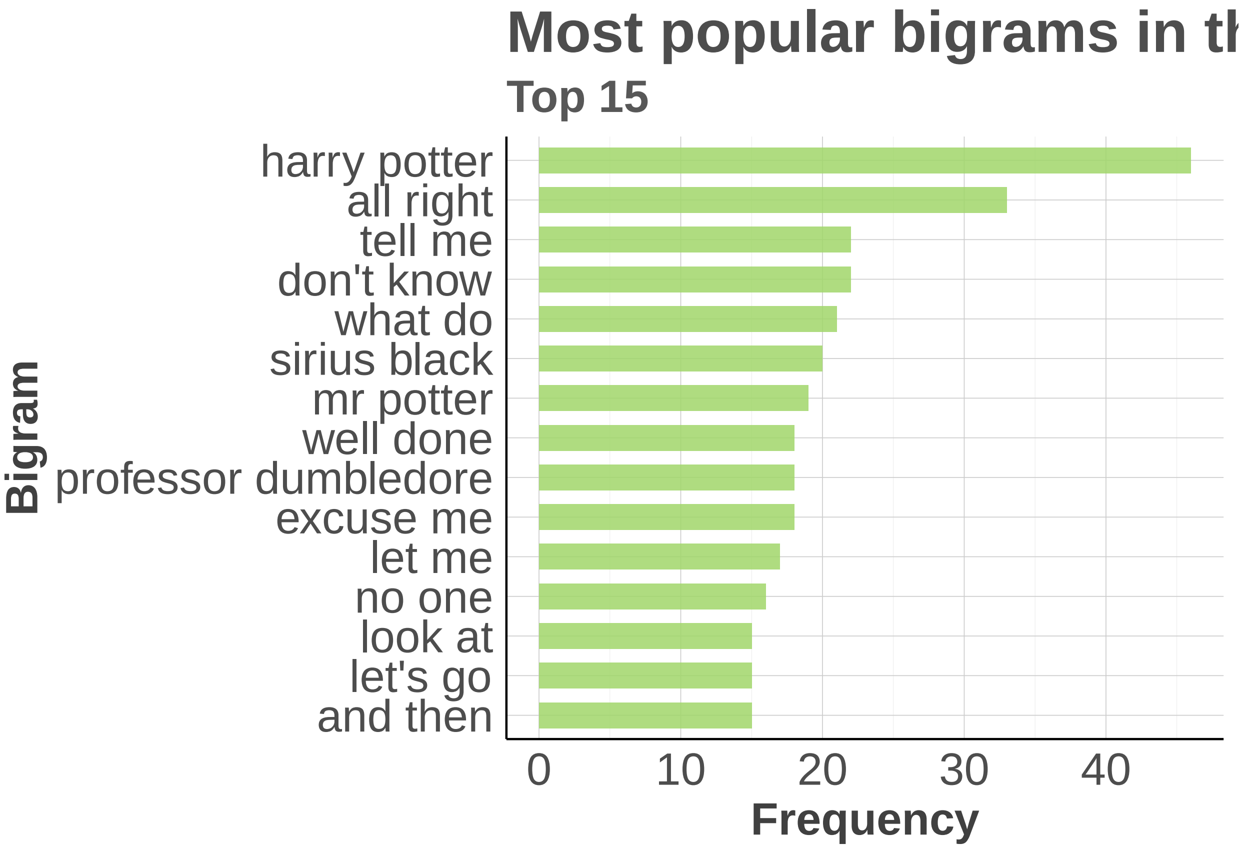


```
tm1 <- Corpus(VectorSource(Script1$Sentence))
tm1 <- tm_map(tm1, content_transformer(tolower))
tm1 <- tm_map(tm1, removeNumbers)
tm1 <- tm_map(tm1, removeWords, stopwords("english"))
tm1 <- tm_map(tm1, removePunctuation)
tm1 <- tm_map(tm1, stripWhitespace)
tdm1 <- TermDocumentMatrix(tm1)

tdm1 <- as.matrix(tdm1)
tdm1 <- sort(rowSums(tdm1), decreasing = T)
tdm1 <- data.frame(Word = names(tdm1), Number = tdm1)
tdm1$Part <- "Sorcerer's Stone"

tm2 <- Corpus(VectorSource(Script2$Sentence))
tm2 <- tm_map(tm2, content_transformer(tolower))
tm2 <- tm_map(tm2, removeNumbers)
tm2 <- tm_map(tm2, removeWords, stopwords("english"))
tm2 <- tm_map(tm2, removePunctuation)
tm2 <- tm_map(tm2, stripWhitespace)
tdm2 <- TermDocumentMatrix(tm2)

tdm2 <- as.matrix(tdm2)
```

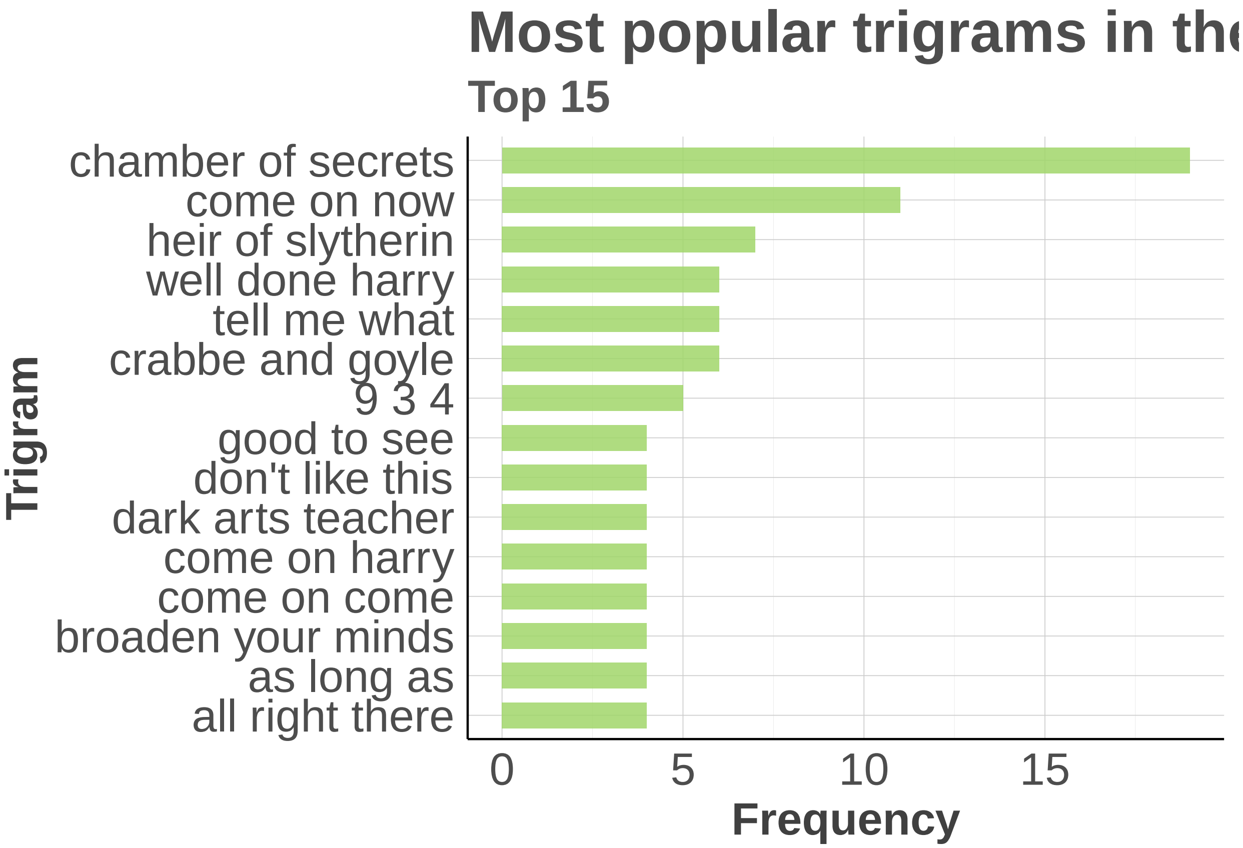
```r
tdm2 <- sort(rowSums(tdm2), decreasing = T)
tdm2 <- data.frame(Word = names(tdm2), Number = tdm2)
tdm2$Part <- "Chamber of Secrets"

tm3 <- Corpus(VectorSource(Script3$Sentence))
tm3 <- tm_map(tm3, content_transformer(tolower))
tm3 <- tm_map(tm3, removeNumbers)
tm3 <- tm_map(tm3, removeWords, stopwords("english"))
tm3 <- tm_map(tm3, removePunctuation)
tm3 <- tm_map(tm3, stripWhitespace)
tdm3 <- TermDocumentMatrix(tm3)

tdm3 <- as.matrix(tdm3)
tdm3 <- sort(rowSums(tdm3), decreasing = T)
tdm3 <- data.frame(Word = names(tdm3), Number = tdm3)
tdm3$Part <- "Prisoner of Azkaban"

tdm_all <- rbind(tdm1, tdm2, tdm3)
tdm_all$Part <- factor(tdm_all$Part, levels=c("Sorcerer's Stone", "Chamber of Secrets", "Prisoner of Az

tdm_all %>%
  filter(Word %in% c("harry", "come", "now", "know", "well", "see", "one", "just", "will",
                     "potter", "think", "right", "can", "professor", "good")) %>%
ggplot(., aes(Word, Number, fill = Part))+
  facet_wrap(.~Part)+
  geom_bar(stat = "identity", width = 0.65, alpha = 0.85)+
  scale_x_discrete(limits = c("good", "professor", "can",  "right", "think", "potter", "will",
                              "just", "one", "see", "well", "know", "now", "come","harry"))+
  scale_fill_manual(values = c("#ffa319", "#767676", "#800000"))+
  coord_flip()+
  labs(title = "Top 15 most popular words in the first 3 movies",
       subtitle = "by part of a movie series",
       x = "Word", y = "Number of sentence per movie")+
  theme_minimal()+
  theme_michau+
  theme(legend.position = "none")
```



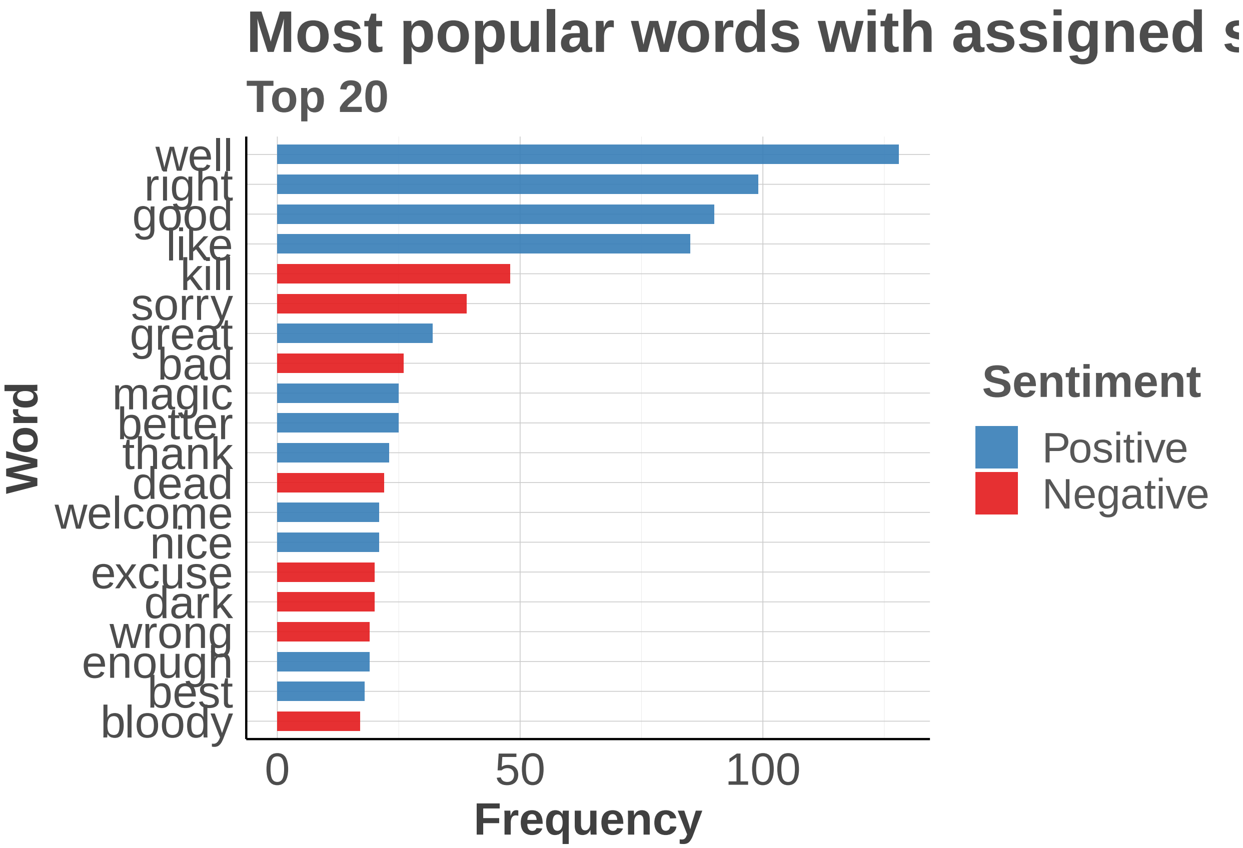
```
Script$Sentence <- as.character(Script$Sentence)

Script %>%
  unnest_tokens(output = word, input = Sentence, token = "ngrams", n = 2) %>%
  filter(is.na(word)==F) %>%
  separate(word, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% c("on", "in", "the", "be", "are", "i", "you", "is", "to", "a", "has", "of", "it",
  filter(!word2 %in% c("on", "in", "the", "be", "are", "i", "you", "is", "to", "a", "has", "of", "it",
  unite(word,word1, word2, sep = " ") %>%
  count(word, sort = T) %>%
  slice(1:15) %>%
ggplot(., aes(reorder(word, +n), n))+
  geom_bar(stat = "identity", width = 0.65, fill = "#a1d76a", alpha = 0.85)+
  coord_flip()+
  labs(title = "Most popular bigrams in the first 3 movies",
       subtitle = "Top 15",
       x = "Bigram", y = "Frequency")+
  theme_minimal()+
  theme_michau
```



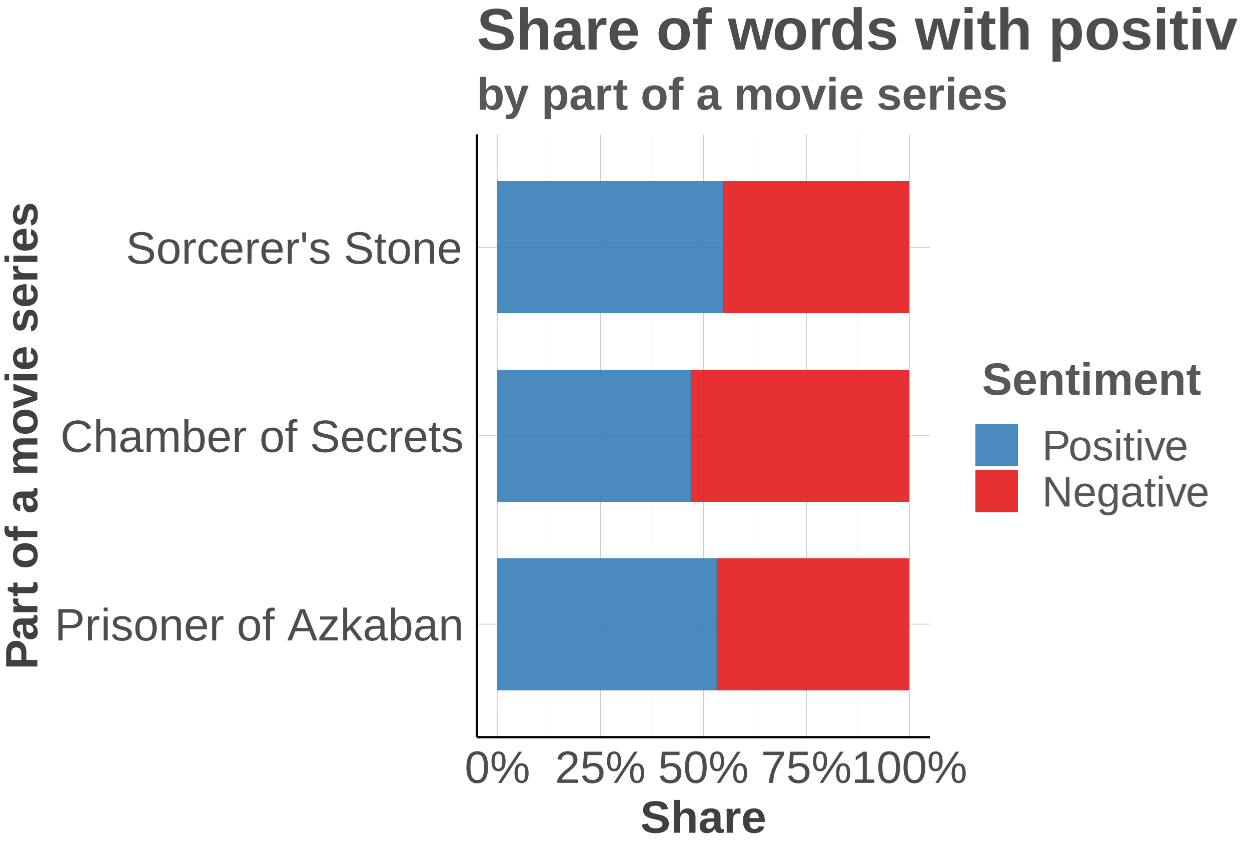
```
Script %>%
  unnest_tokens(output = word, input = Sentence, token = "ngrams", n = 3) %>%
  filter(is.na(word)==F) %>%
  separate(word, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% c("on", "in", "the", "be", "are", "i", "you", "is", "to", "a", "has", "of", "it",
  filter(!word2 %in% c("you", "we", "the")) %>%
  filter(!word3 %in% c("on", "in", "the", "be", "are", "i", "you", "is", "to", "a", "has", "of", "it",
  unite(word,word1, word2, word3, sep = " ") %>%
  count(word, sort = T) %>%
  slice(1:15) %>%
ggplot(., aes(reorder(word, +n), n))+
  geom_bar(stat = "identity", width = 0.65, fill = "#a1d76a", alpha = 0.85)+
  coord_flip()+
  labs(title = "Most popular trigrams in the first 3 movies",
       subtitle = "Top 15",
       x = "Trigram", y = "Frequency")+
  theme_minimal()+
  theme_michau
```



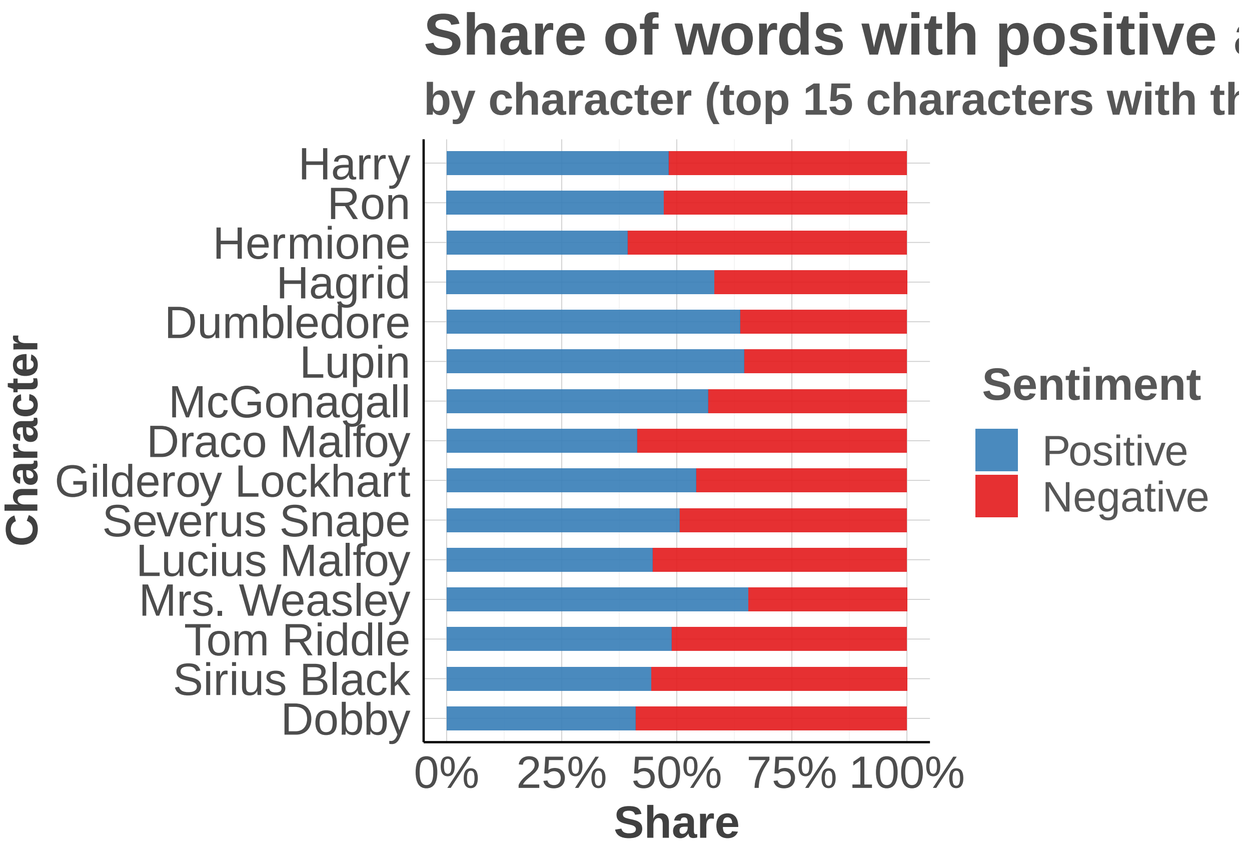
```
Sentiment <- Script %>%
  unnest_tokens(output = word, input = Sentence) %>%
  left_join(Bing, "word") %>%
  filter(is.na(sentiment)==F)

Sentiment %>%
  group_by(word, sentiment) %>%
  summarise(count = n(), .groups = 'drop') %>%
  arrange(desc(count)) %>%
  slice(1:20) %>%
ggplot(., aes(reorder(word, +count), count, fill = sentiment))+
  geom_bar(stat = "identity", width = 0.65, alpha = 0.9)+
  scale_fill_brewer(palette = "Set1")+
  coord_flip()+
  labs(title = "Most popular words with assigned sentiment",
       subtitle = "Top 20",
       x = "Word", y = "Frequency", fill = "Sentiment")+
  guides(fill = guide_legend(reverse = T))+
  theme_minimal()+
  theme_michau+
  theme(legend.title.align = 0.5, legend.position = "right", legend.direction = "vertical")
```



```
Sentiment %>%
  group_by(Part, sentiment) %>%
  summarise(count = n(), .groups = 'drop') %>%
ggplot(., aes(Part, count, fill = sentiment))+
  geom_bar(stat = "identity", position = "fill", width = 0.7, alpha = 0.9)+
  scale_fill_brewer(palette = "Set1")+
  scale_y_continuous(labels = scales::percent)+
  coord_flip()+
  labs(title = "Share of words with positive and negative sentiment",
       subtitle = "by part of a movie series", fill = "Sentiment",
       x = "Part of a movie series", y = "Share")+
  guides(fill = guide_legend(reverse = T))+
  theme_minimal()+
  theme_michau+
  theme(legend.title.align = 0.5, legend.position = "right", legend.direction = "vertical")
```



```
Sentiment %>%
  filter(Character %in% c("Harry", "Ron", "Hermione", "Hagrid", "Dumbledore", "Lupin", "McGonagall", "D
                          "Severus Snape", "Lucius Malfoy", "Mrs. Weasley", "Tom Riddle", "Sirius Black
  group_by(Character, sentiment) %>%
  summarise(count = n(), .groups = 'drop') %>%
ggplot(., aes(Character, count, fill = sentiment))+
  geom_bar(stat = "identity", position = "fill", width = 0.6, alpha = 0.9)+
  scale_x_discrete(limits = c("Dobby", "Sirius Black", "Tom Riddle", "Mrs. Weasley", "Lucius Malfoy", "S
                              "McGonagall", "Lupin", "Dumbledore", "Hagrid",  "Hermione", "Ron", "Harry
  scale_fill_brewer(palette = "Set1")+
  scale_y_continuous(labels = scales::percent)+
  coord_flip()+
  labs(title = "Share of words with positive and negative sentiment",
       subtitle = "by character (top 15 characters with the most sentences)", fill = "Sentiment",
       x = "Character", y = "Share")+
  guides(fill = guide_legend(reverse = T))+
  theme_minimal()+
  theme_michau+
  theme(legend.title.align = 0.5, legend.position = "right", legend.direction = "vertical")
```



Share of words with positive

by character (top 15 characters with th