EU PDF Document Report

Lilian

16-August 2023

```r
#import pdf file
EU_report <- c("C:/Users/LILIAN/Desktop/R Code/EUREPORT.pdf")

#view
EU_report
```

```
## [1] "C:/Users/LILIAN/Desktop/R Code/EUREPORT.pdf"
```

```r
#EXPLORE AND MANIPULATE THE PDF FILE

#capture meta data on pdf file (This gives all the info about the file, the version, how many pages, if
info <- pdf_info("C:/Users/LILIAN/Desktop/R Code/EUREPORT.pdf")


#capture the text(This gives you information about each page of the file)
text <- pdf_text("C:/Users/LILIAN/Desktop/R Code/EUREPORT.pdf")


#to check the datatype
typeof(EU_report1)
```

```
## [1] "list"
```

```r
#we have to change the datatype to a vector because the library tidytext works with a vector
EU_report1 <- unlist(EU_report1)


#tidytext convert it to a dataframe format before converting to tokens
EU <- data_frame(line = 1:94, text = EU_report1)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```
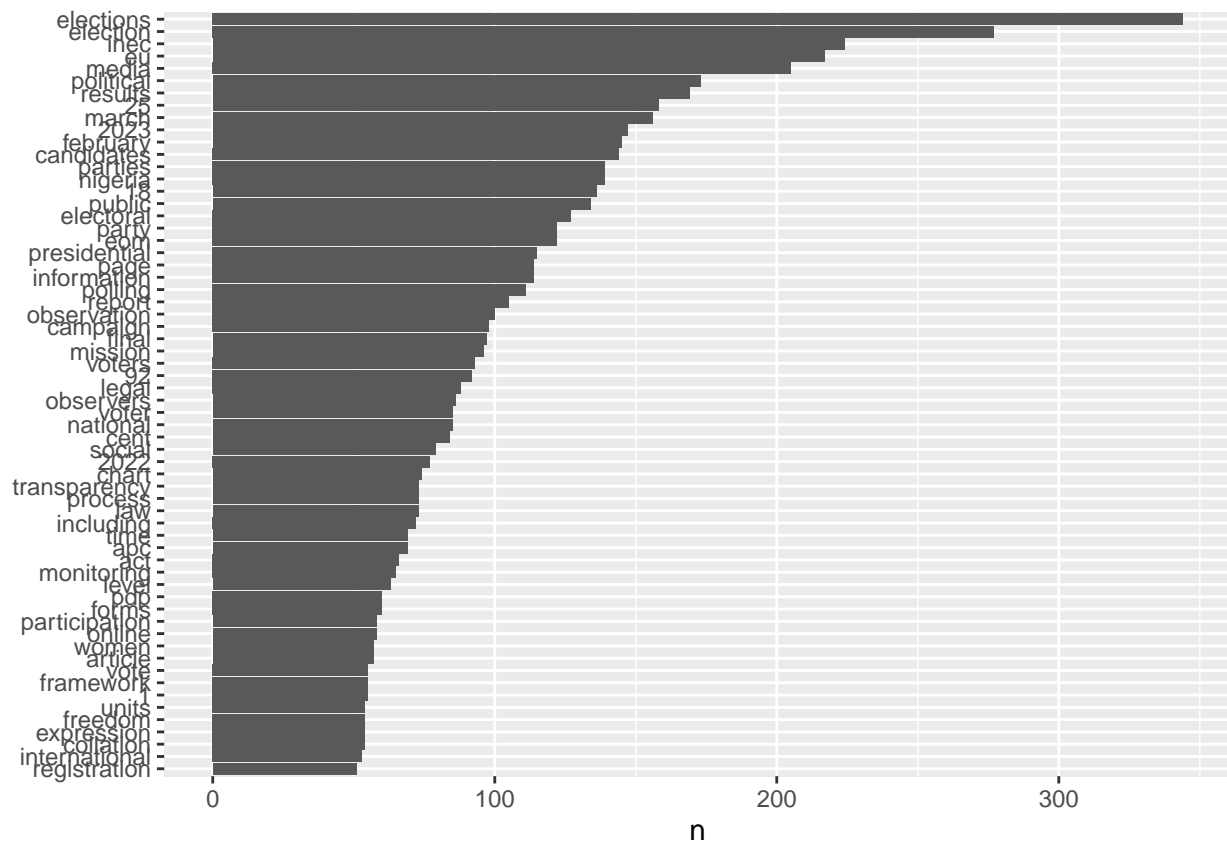
```r
view(EU)
```

```r
#remove stop word
EU_pdf1 <- EU_pdf %>%
  anti_join(stop_words)
```

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

```
#plotting a histogram of a word that appeared more than 30
EU_pdf1 %>%
  count(word, sort = TRUE) %>%
  filter(n > 50) %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```



```
#creating a word cloud
EU_pdf1 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 120, colors = brewer.pal(6, "Dark2")))
```

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

```
## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): political could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): nigeria could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): february could not be fit on page. It will not be plotted.

## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): elections could not be fit on page. It will not be plotted.

## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): election could not be fit on page. It will not be plotted.

## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): observation could not be fit on page. It will not be plotted.

## Warning in wordcloud(word, n, max.words = 120, colors = brewer.pal(6,
## "Dark2")): polling could not be fit on page. It will not be plotted.
```