Lab sheet 7b: Simple sentiment analysis

Load necessary packages

```
library(tidyverse)
library(tidytext)
library(rvest)
library(textdata)
library(wordcloud)
library(RColorBrewer)
library(wordcloud2)
```

Use the rvest package to scrape and extract text from a website

```
url <- 'https://thewire.in/world/us-debt-ceiling-explained'
news <- read_html(url)
```

Get texts

```
text <- tibble(
  news %>%
    html_elements('p') %>% #it selects all paragraph elements (<p>) from the HTML content
    html_text()
) %>%
  rename( 'text' = 1) # rename the colum to "text"
```

Get sentiments

```
sentiments <- get_sentiments('nrc')
```

Seperate all texts

```
tokens <- text %>%
  unnest_tokens(input = text,output = word)%>%
  filter(!grepl('[0-9]', word)) #remove numbers
```

Remove stop-words and count each words

```
word_freq <- tokens%>%anti_join(stop_words) %>%
  count(word, sort=TRUE)
```

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

wordcloud plots

```
set.seed(1234)
wordcloud(words = word_freq$word,
          freq = word_freq$n,
          min.freq = 1,
          max.words=150,
          random.order=FALSE,
          rot.per=0.40,
          colors=brewer.pal(9, "Dark2")
          )
```


```
wordcloud2(
  data = word_freq,
  #size = 1.0,   # Size of words
  color = "random-dark",  # Color scheme
  backgroundColor = "white",  # Background color
  fontFamily = "Arial",  # Font family
  minRotation = -pi/4,  # Minimum rotation angle
  maxRotation = pi/4   # Maximum rotation angle
)
```

Word freqency count and add sentiment score to each words

```
freq_count <- tokens %>%
  inner_join(sentiments, by='word', multiple = "all") %>%
  count(sentiment, sort = TRUE)
```

Plot the result

```
gg <- freq_count %>%
  ggplot( aes(x= sentiment, y= n,
              fill= sentiment) ) +
  geom_col(show.legend = F) +
  geom_text(aes(x= sentiment, y= n, label=n), size=8) +
  labs( x= 'Sentiment',
        y= 'Frequency',) +
  ggtitle('All sentiments with frequencies')+
  theme(plot.subtitle =  element_text(color = "gray", face = "italic")) +
  theme_light()
gg
```