4 Chan

Webscraper

Anon

Load Libraries

```
library("rvest")
library("tidyverse")
library("tidytext")
library("ggplot2")
library("wordcloud")
```

```
#Page 1 Scrape
pol_threads1 <- read_html("https://boards.4channel.org/pol/") %>%
   html_elements("blockquote.postMessage") %>%
   html_text()

#Page 2 Scrape
pol_threads2 <- read_html("https://boards.4channel.org/pol/2") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 3 Scrape
pol_threads3 <- read_html("https://boards.4channel.org/pol/3") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 4 Scrape
pol_threads4 <- read_html("https://boards.4channel.org/pol/4") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 5 Scrape
pol_threads5 <- read_html("https://boards.4channel.org/pol/5") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 6 Scrape
pol_threads6 <- read_html("https://boards.4channel.org/pol/6") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 7 Scrape
pol_threads7 <- read_html("https://boards.4channel.org/pol/7") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()
```

```r
#Page 8 Scrape
pol_threads8 <- read_html("https://boards.4channel.org/pol/8") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 9 Scrape
pol_threads9 <- read_html("https://boards.4channel.org/pol/9") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()

#Page 10 Scrape
pol_threads10 <- read_html("https://boards.4channel.org/pol/10") %>%
  html_elements("blockquote.postMessage") %>%
  html_text()
```

```r
#tibble makes a table out of data
df_pol <-      c(pol_threads1,
                 pol_threads2,
                 pol_threads3,
                 pol_threads4,
                 pol_threads5,
                 pol_threads6,
                 pol_threads7,
                 pol_threads8,
                 pol_threads9,
                 pol_threads10)

pol_table <- tibble(txt = df_pol)
pol_table
```

```
## # A tibble: 1,068 x 1
##    txt
##    <chr>
##  1 This board is for the discussion of news, world events, political issues, an~
##  2 Check the catalog before posting a new thread!Reply to existing threads abou~
##  3 Are we actually going to see bodies in the street this time around?
##  4 >>438851797so 96% of people under 80 will survive opposed to 98%
##  5 >>438856541What an optimistic view. I just don't see it. They already have a~
##  6 >>438854194God I hope this is true.
##  7 >>438851797
##  8 >>438856234yet people still hate nazi's... i don't get this world.
##  9 RAPED TO DEATHBY ANIMALSTHAT IS THE ENDOF SODOMYTHIS IS THE CURSETHEY WILL B~
## 10 But why?https://youtu.be/zWJJ5XDEpOsHe already made his video unavailable.
## # i 1,058 more rows
```

```r
tidy_pol <- pol_table %>%
  unnest_tokens(word, txt, format = "text")

tidy_pol_fixed <- tidy_pol %>%
  filter(!word %in% stop_words$word
         & !word == "fucking"
         & !word == "https"
         & !word == "shit"
         & !is.numeric(word))
```

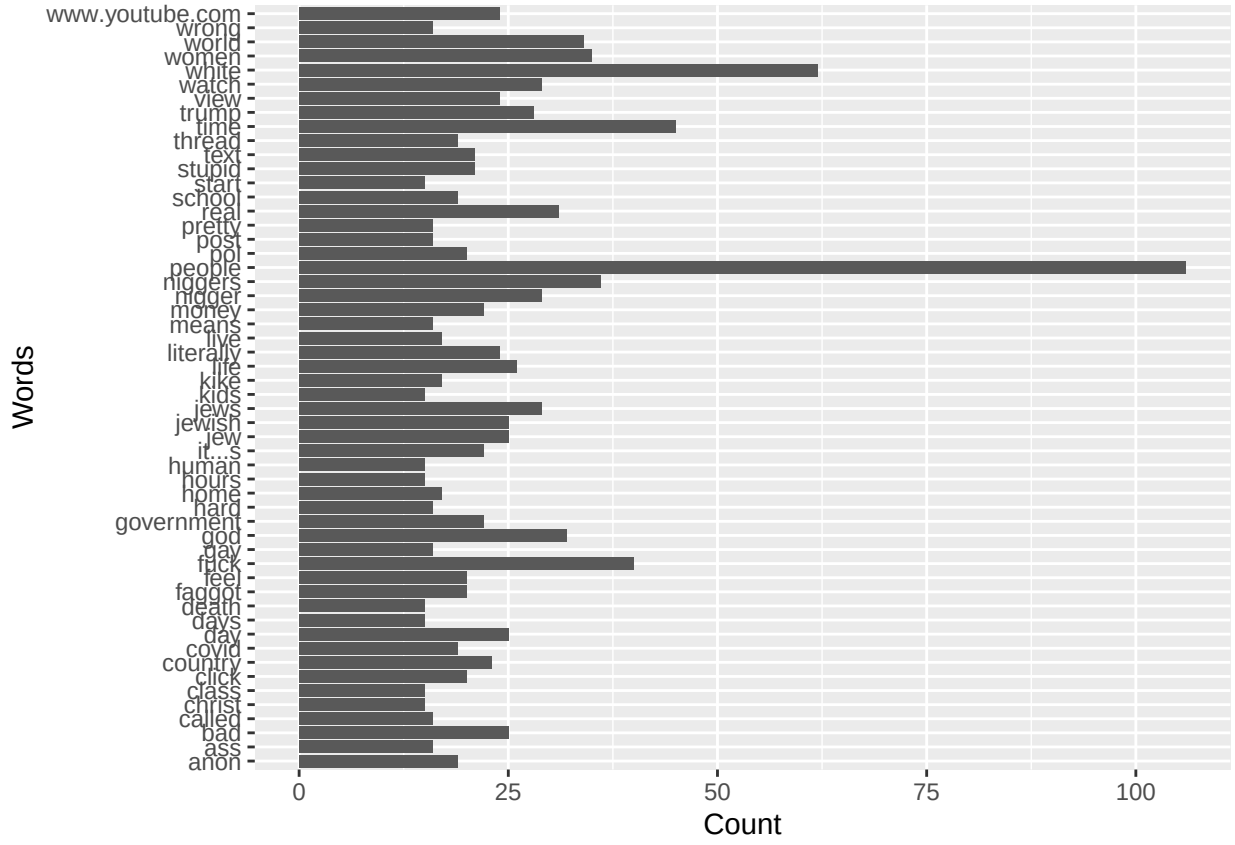
```r
tidy_pol_fixed2 <- tidy_pol_fixed %>%
  count(word, sort = TRUE) %>%
  print(n = 50)
```

```
## # A tibble: 6,540 x 2
##    word                n
##    <chr>           <int>
##  1 people            106
##  2 white              62
##  3 time               45
##  4 fuck               40
##  5 niggers            36
##  6 women              35
##  7 world              34
##  8 god                32
##  9 real               31
## 10 jews               29
## 11 nigger             29
## 12 watch              29
## 13 trump              28
## 14 life               26
## 15 bad                25
## 16 day                25
## 17 jew                25
## 18 jewish             25
## 19 literally          24
## 20 view               24
## 21 www.youtube.com    24
## 22 country            23
## 23 government         22
## 24 it's               22
## 25 money              22
## 26 stupid             21
## 27 text               21
## 28 click              20
## 29 faggot             20
## 30 feel               20
## 31 pol                20
## 32 anon               19
## 33 covid              19
## 34 school             19
## 35 thread             19
## 36 home               17
## 37 kike               17
## 38 live               17
## 39 ass                16
## 40 called             16
## 41 gay                16
## 42 hard               16
## 43 means              16
## 44 post               16
## 45 pretty             16
## 46 wrong              16
```

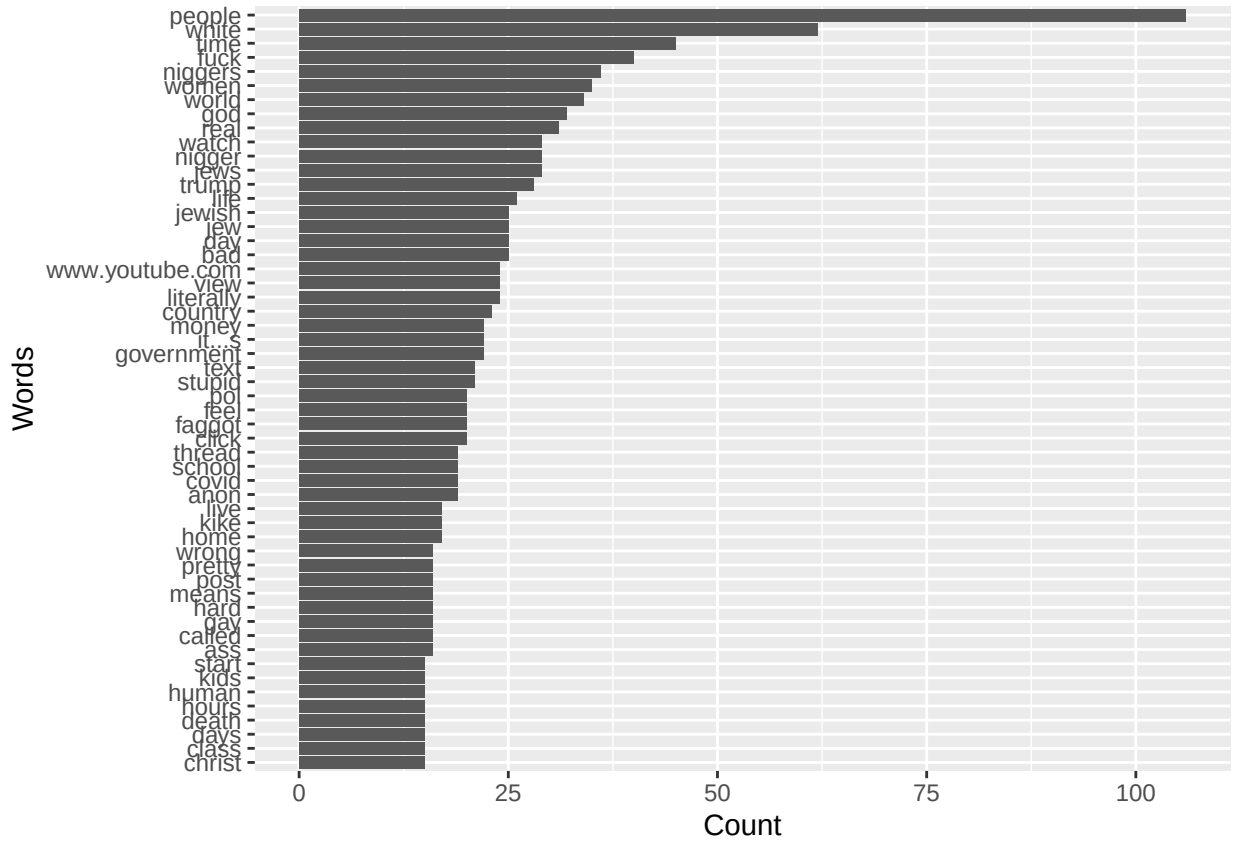
```
## 47 christ              15
## 48 class               15
## 49 days                15
## 50 death               15
## # i 6,490 more rows
```

Time to Visualize the word data.

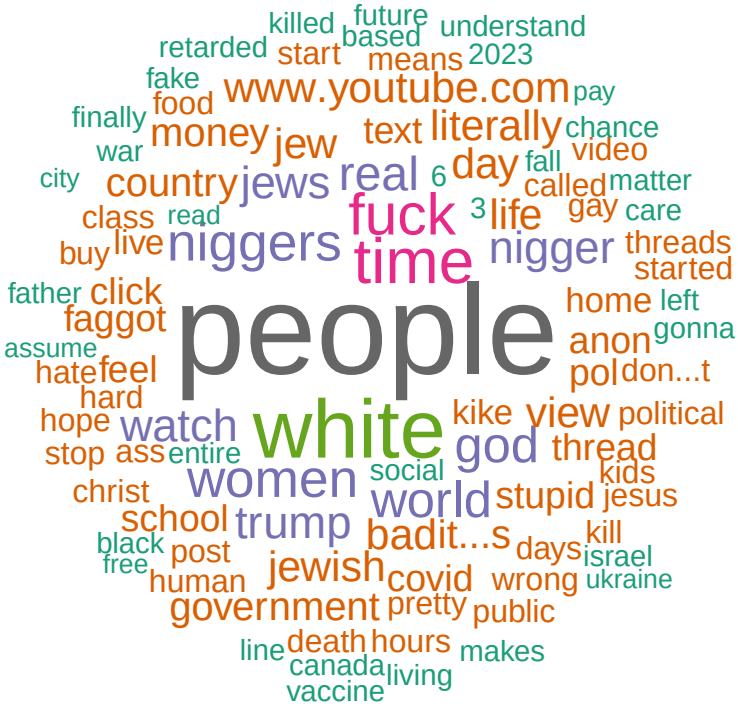
```
tidy_pol_fixed2 %>%
  top_n(50) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab("Words") +
  ylab("Count") +
  coord_flip()
```




```
tidy_pol_fixed2 %>%
  top_n(50) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab("Words") +
  ylab("Count") +
  coord_flip()
```




```
tidy_pol_fixed2 %>%
  with(wordcloud(word, n, max.words = 100, random.order = FALSE, rot.per = 0.0,
                                colors = brewer.pal(8, "Dark2")))
```



Save the Data. Make sure to change the date when saving to not overwrite the old data

write.csv(tidy_pol_fixed, "~/Documents/Stats/4Chan Scraper/Aug-22-2023-1116h.csv", row.names=FALSE)