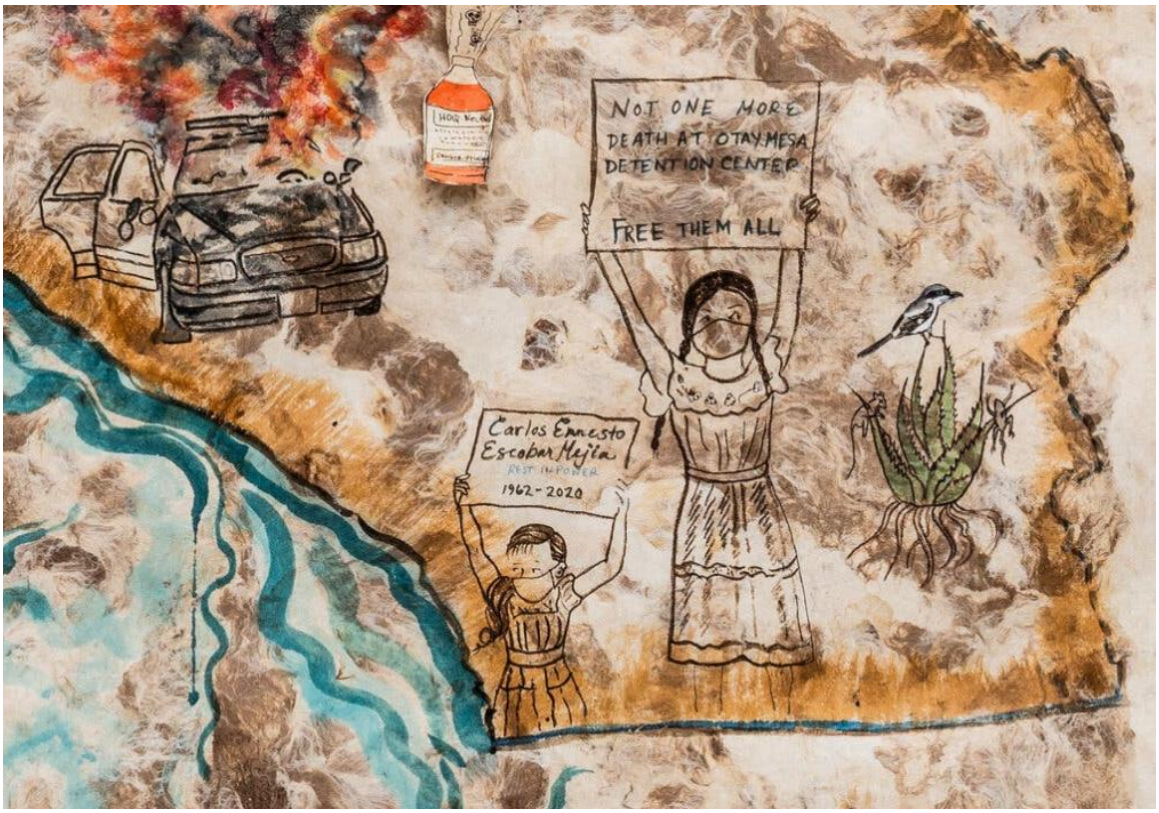
TEXT MINING WITH R

DR. ALVIN ANG



CONTENTS

I. STEP 1: INSTALL TIDYVERSE

File: https://www.alvinang.sg/s/Text-Mining-with-R-by-Dr-Alvin-Ang.R

```
#Step 0: Getting the Files
#https://www.alvinang.sg/s/eisenhower.txt
```

```
#Step 1: Installing Tidyverse into R
installed.packages("tidyverse", dependencies = TRUE)
```

- https://www.tidyverse.org/

A. INSTALLIING TIDYVERSE INTO LINUX MINT

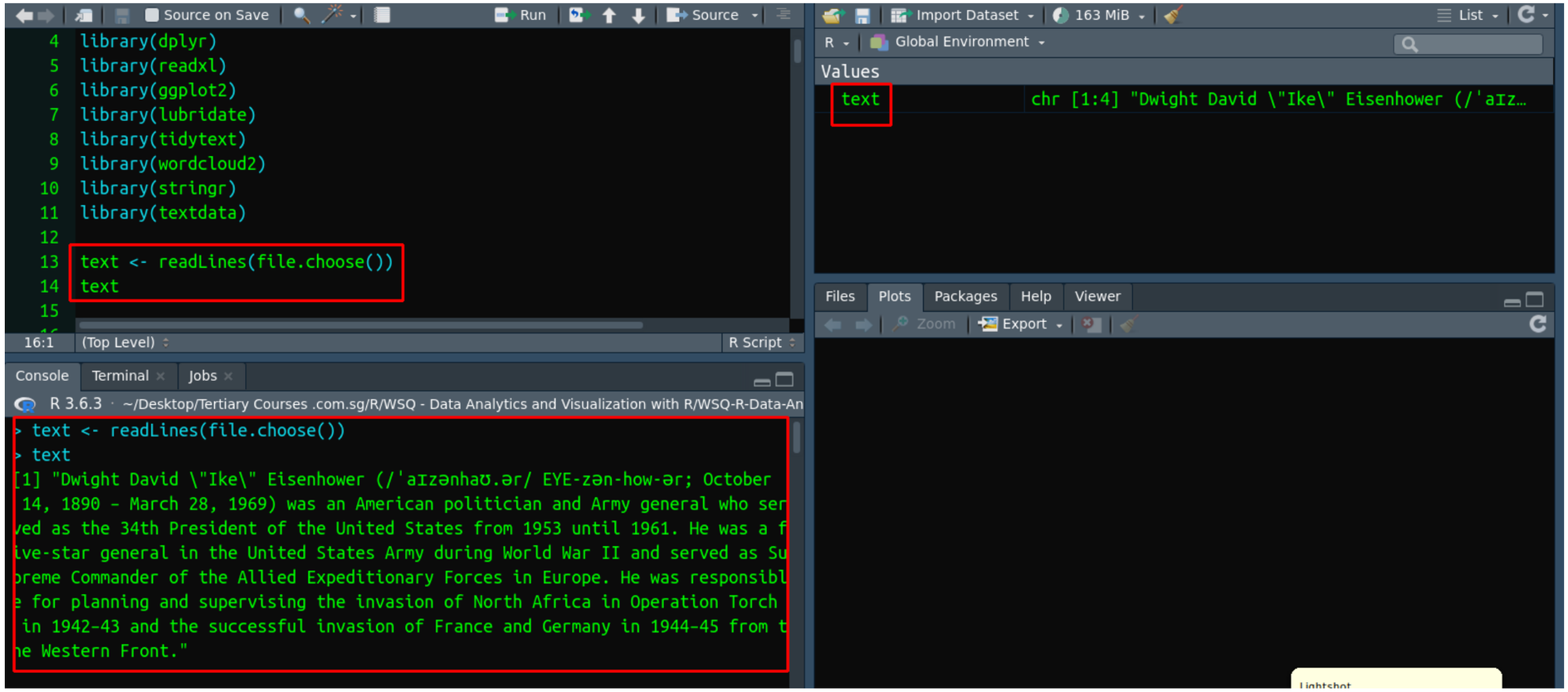
- You most probably have no issues installing Tidyverse into R using Windows.
- But Linux Mint is tough.
- Do the following:
    - sudo apt install g++
    - sudo apt-get update
    - sudo apt-get install libcurl4-openssl-dev
    - sudo apt-get install r-base-dev.
    - reboot your laptop
    - reinstall tidyverse:
        - install.packages("tidyverse", dependencies=TRUE)
    - sudo apt install libssl-dev libxml2-dev

II. STEP 2: LOAD PACKAGES

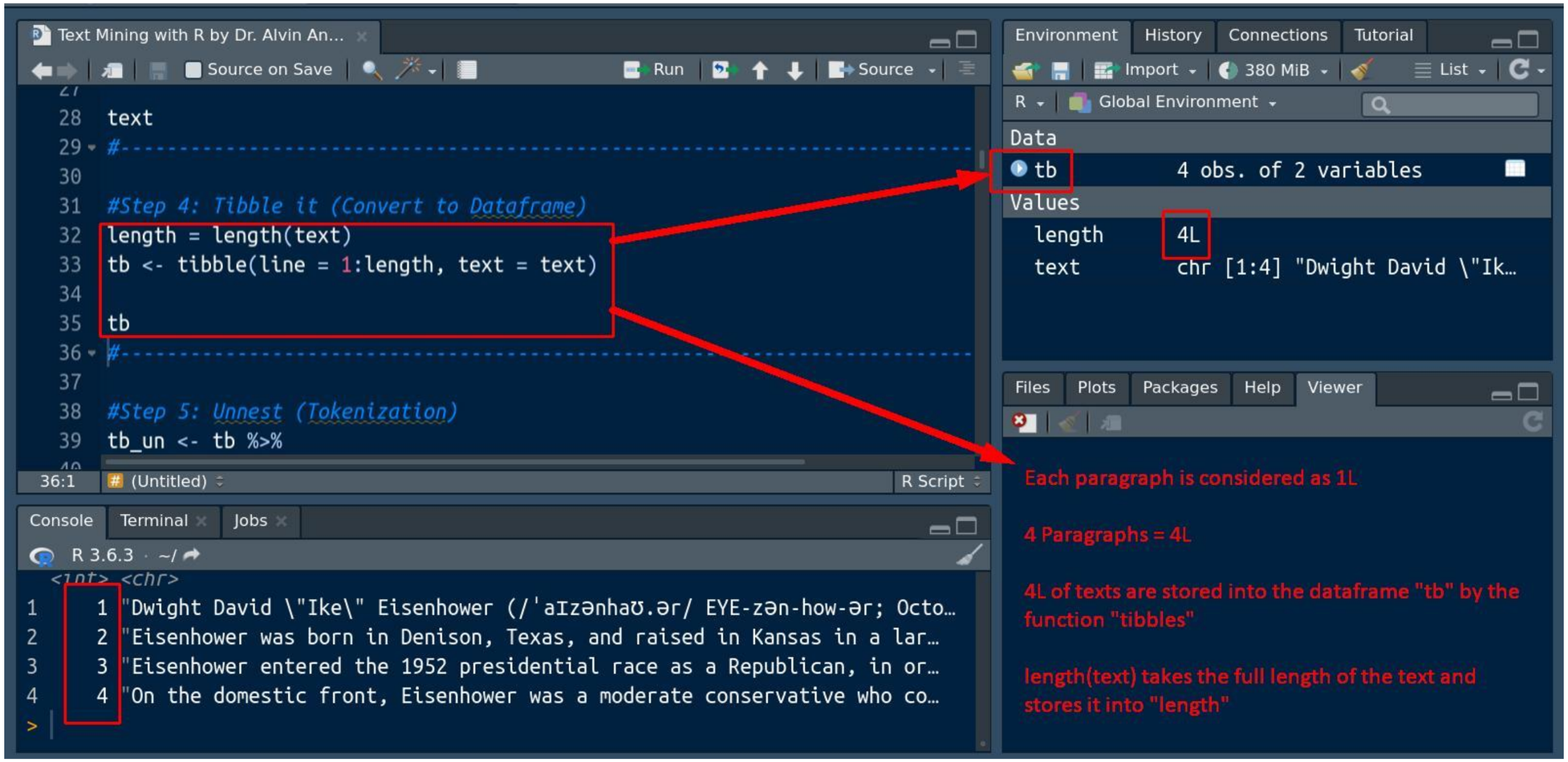
```
#Step 2: Load Packages
library(tidyverse)
library(tibble)
library(tidyr)
library(dplyr)
library(readxl)
library(ggplot2)
library(ggthemes)
library(lubridate)
library(tidytext)
library(wordcloud2)
library(stringr)
library(textdata)
```

File can be found here: https://www.alvinang.sg/s/eisenhower.txt

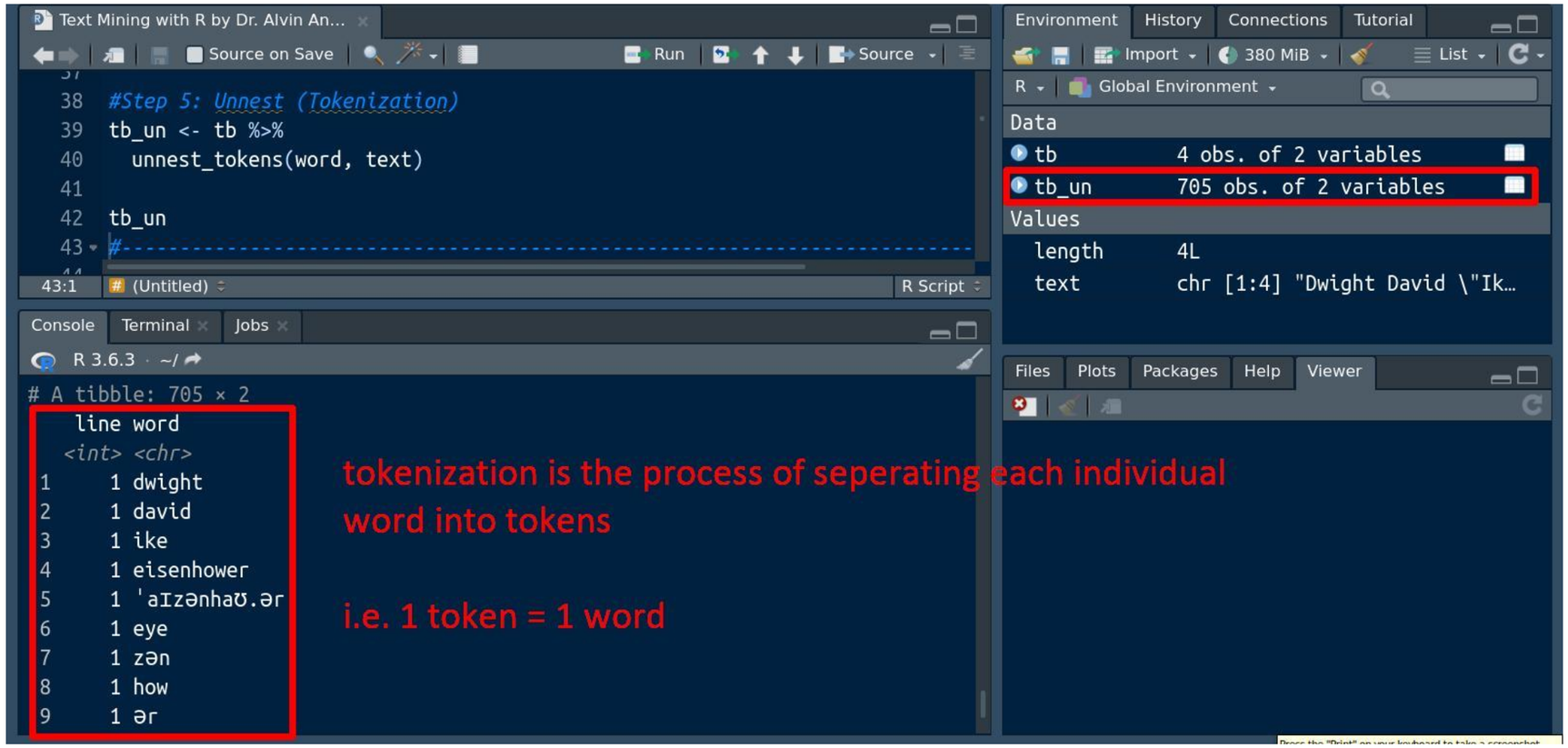
```
#Step 3: Import Text File
text <- readLines(file.choose())

text
```



```
#Step 4: Tibble it (Convert to Dataframe)
length = length(text)
tb <- tibble(line = 1:length, text = text)

tb
```



```
#Step 5: Unnest (Tokenization)
tb_un <- tb %>%
  unnest_tokens(word, text)

tb_un
```



Stop Words

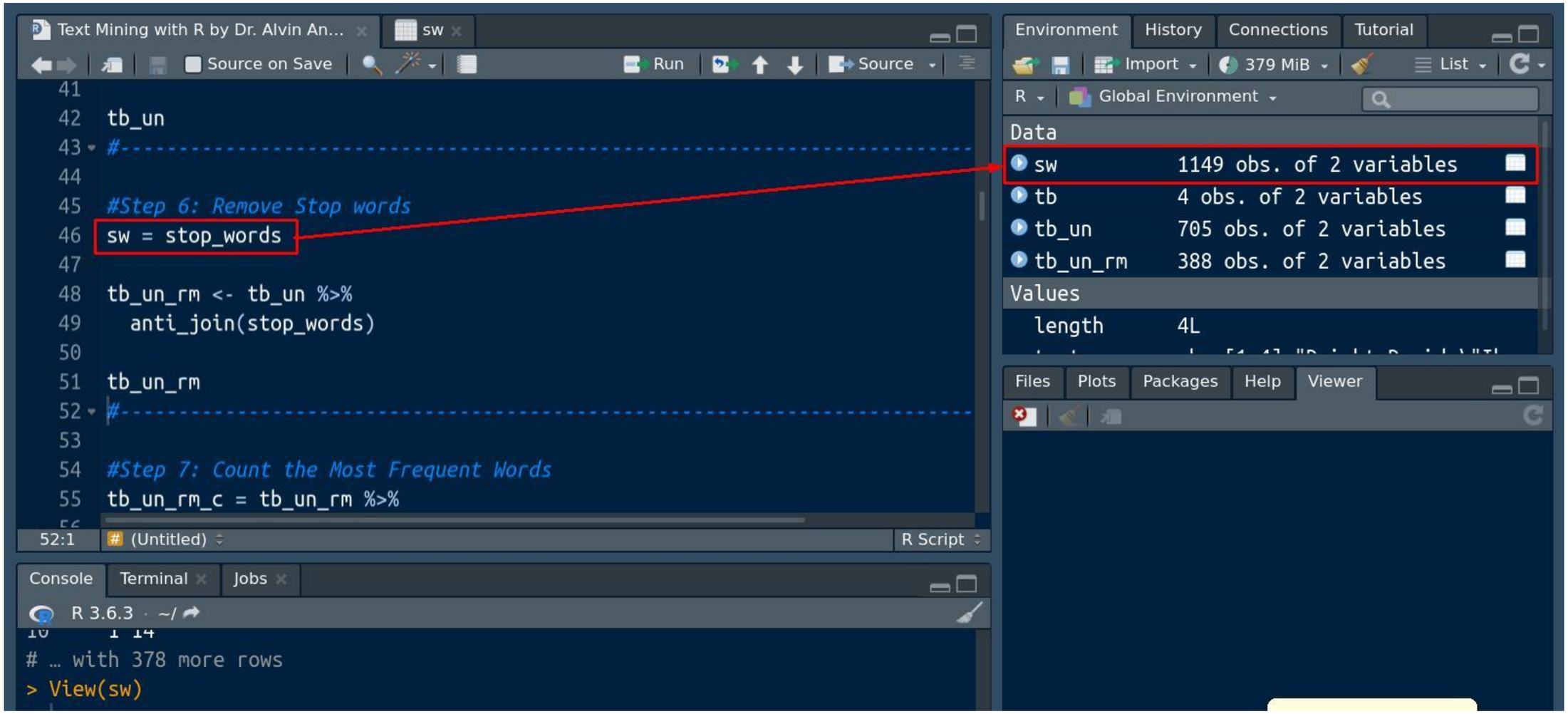
These words include:

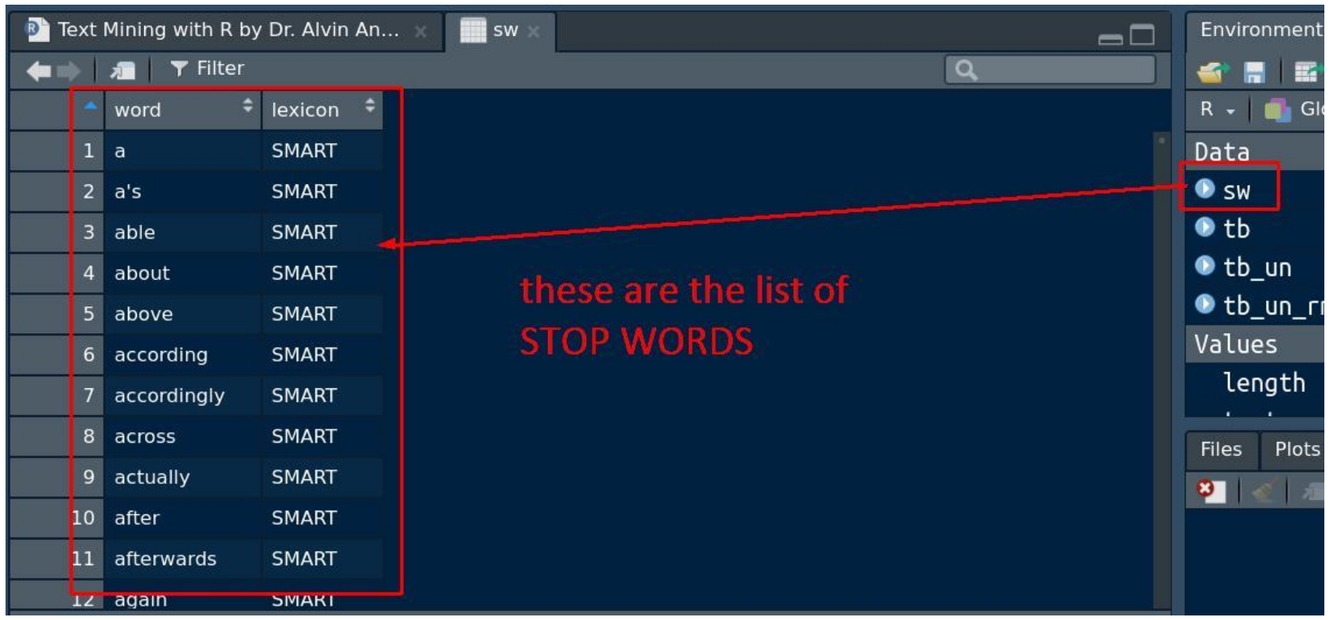
- a
- I
- the
- in
- of
- for
- at
- to
- on
- with
- from

Stop words are useless words when it comes to text analysis, because they don't have meaning. We will remove them.

```
#Step 6: Remove Stop words
sw = stop_words

tb_un_rm <- tb_un %>%
  anti_join(stop_words)

tb_un_rm
```

A. SHOWING THE LIST OF STOP WORDS

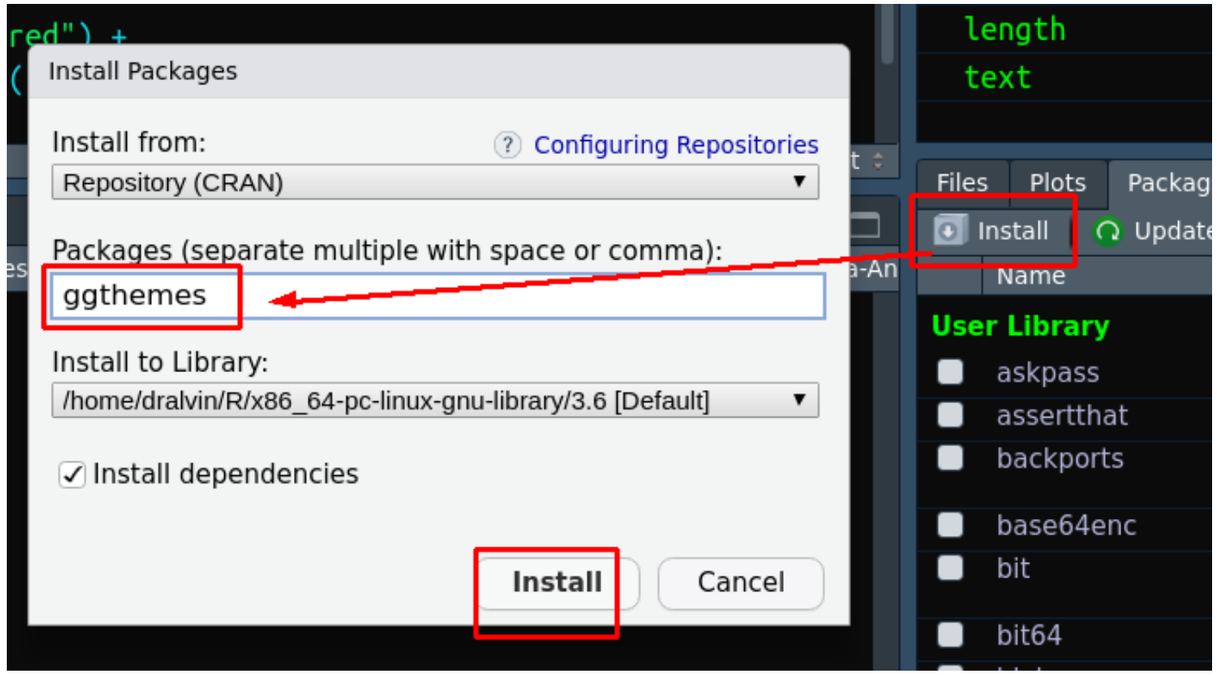


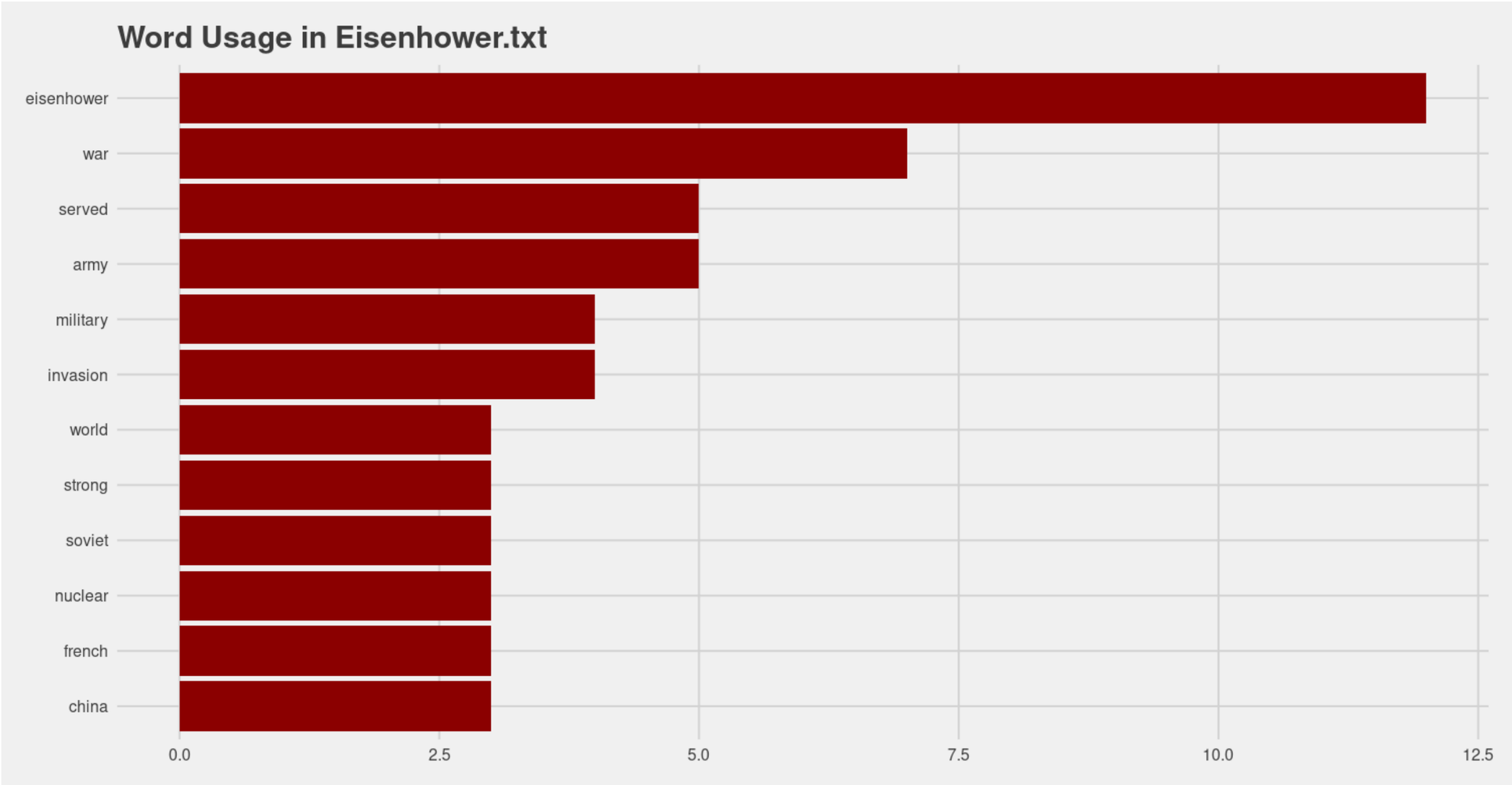


```
#Step 7: Count the Most Frequent Words
tb_un_rm_c = tb_un_rm %>%
  count(word, sort = TRUE)

tb_un_rm_c
```

```
> tidy_df_rm %>%
+   count(word, sort = TRUE)
# A tibble: 300 × 2
   word          n
   <chr>     <int>
 1 eisenhower   12
 2 war           7
 3 army          5
 4 served        5
 5 invasion      4
 6 military      4
 7 china         3
 8 french        3
 9 nuclear       3
10 soviet        3
# … with 290 more rows
```

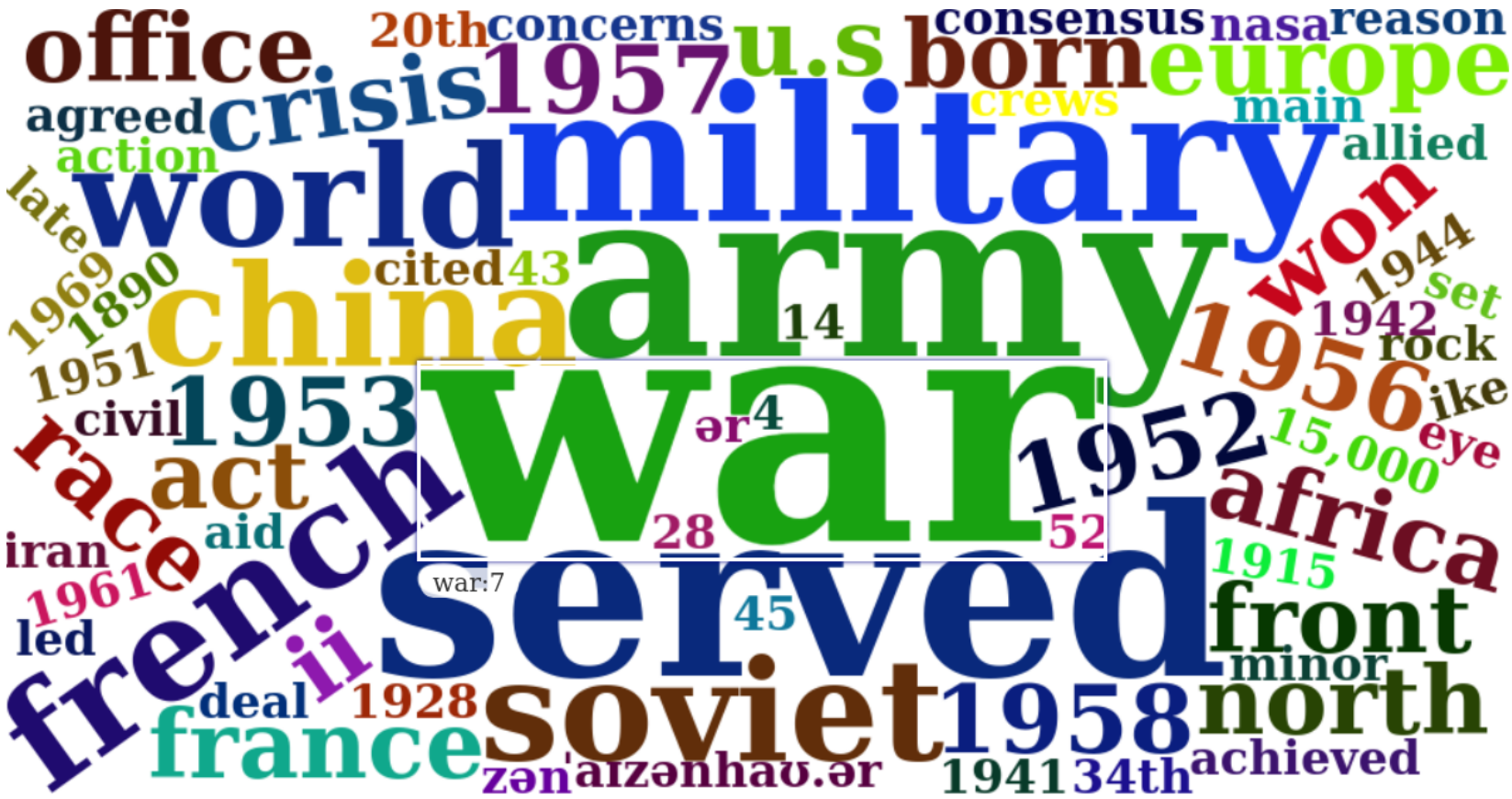
```r
#Step 8: Plotting the Most Frequent Words
tb_un_rm_c_plt = tb_un_rm_c %>%
  filter(n>2) %>%
  mutate(word = reorder(word, n)) %>%

  ggplot(aes(word, n)) + geom_col(fill = "darkred") + theme_fivethirtyeight() +
  xlab(NULL) + ylab("Word Count") + coord_flip() + ggtitle("Word Usage in Eisenhower.txt")

tb_un_rm_c_plt
```





.

```
#Step 9: Creating Word Cloud
tb_un_rm_cld = tb_un_rm %>%
  count(word, sort=TRUE) %>%
  wordcloud2(word,size=2)

tb_un_rm_cld
```



The bigger the word, the greater the word count.

X. STEP 10: ANALYZE TEXT SENTIMENTS

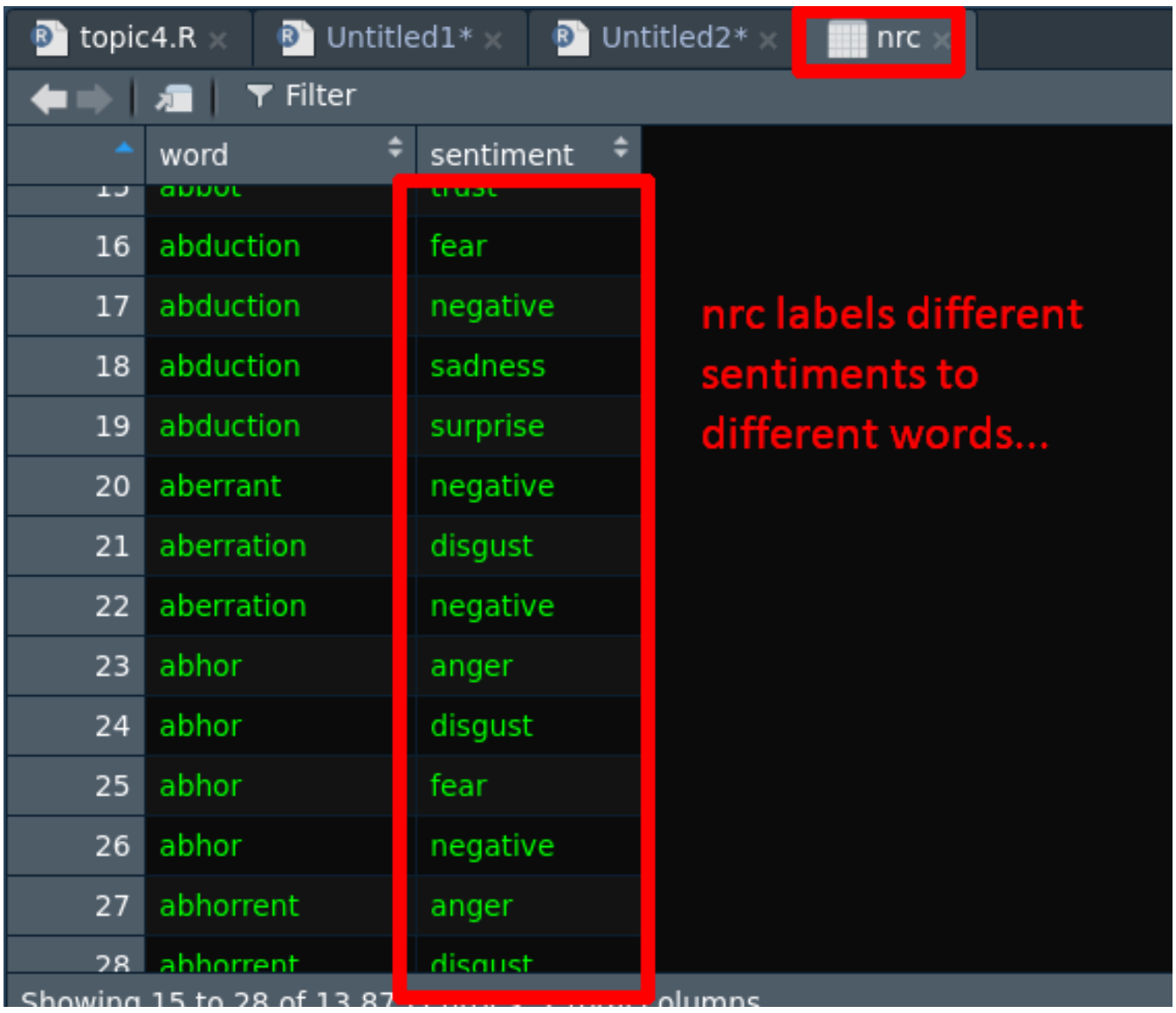
- Sentiment lexicons give emotions to a given text.
- There are three lexicons that can be used:

A. NRC

1. GET SENTIMENTS

```
#Step 10: Analyze Text Sentiments
#10a) NRC
# nrc categorizes words as POSITIVE / NEGATIVE / ANGER / ANTICIPATION / DISGUST
# FEAR / JOY / SADNESS / SURPRISE AND TRUST

#10a)(i) Get Sentiments
nrc = get_sentiments("nrc")
```

```
Selection: 1   nrc is a big library... u have to wait long time to load it...
trying URL 'http://saifmohammad.com/WebDocs/NRC-Emotion-Lexicon.zip'
Content type 'application/zip' length 24436570 bytes (23.3 MB)
==================================
```



2. FILTER SENTIMENTS

```
#10a)(ii) Filter Sentiments
nrc_sentiment <- get_sentiments("nrc") %>%
  filter(sentiment == "fear")
```

3. INNER JOIN NRC "FEAR" TO TEXT FILE + DO WORD COUNT

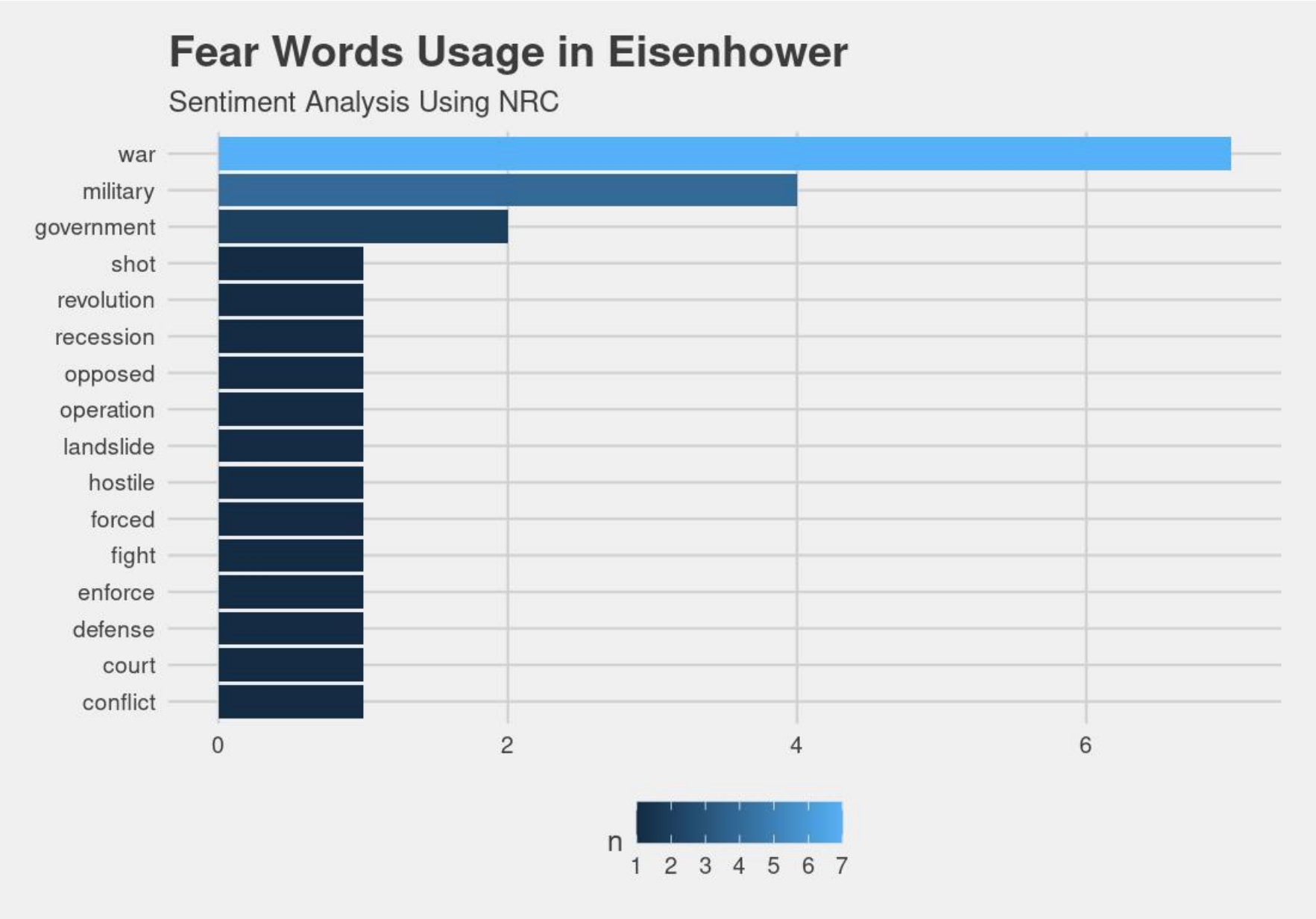
```
#10a)(iii) Inner Join Sentiments to Text File + Do Word Count
tb_un_rm_nrc = tb_un_rm %>%
  inner_join(nrc_sentiment) %>%
  count(word, sort = TRUE)

tb_un_rm_nrc
```

```
   word            n
   <chr>       <int>
 1 war             7
 2 military        4
 3 government      2
 4 conflict        1
 5 court           1
 6 defense         1
 7 enforce         1
 8 fight           1
 9 forced          1
10 hostile         1
11 landslide       1
12 operation       1
13 opposed         1
14 recession       1
15 revolution      1
16 shot            1
>
```

From the Eisenhower.txt, we see that 16 words affiliate with "fear", the top word striking fear is "war".

4. VISUALIZE

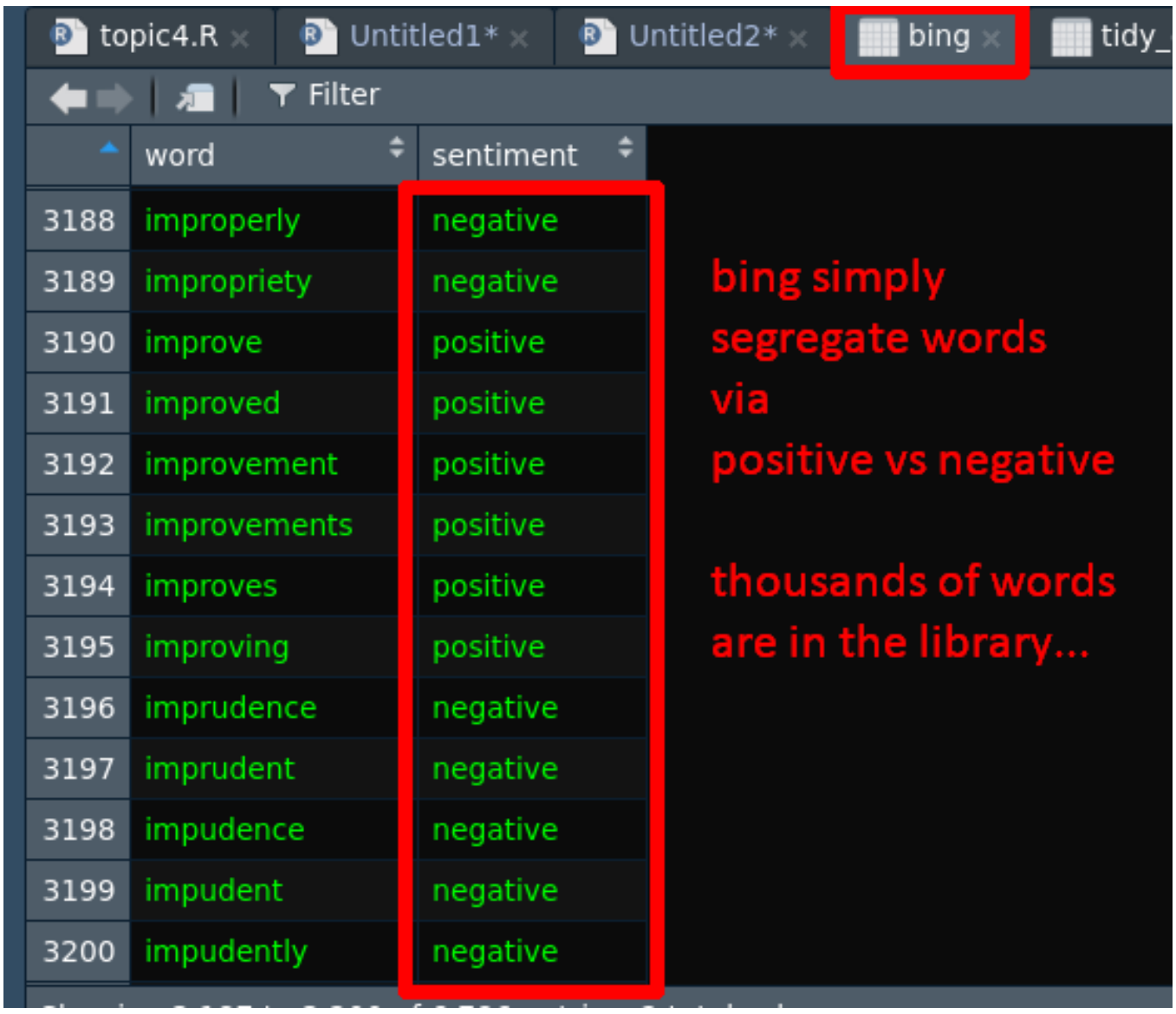
```
#10a)(iv) Visualize
tb_un_rm_nrc %>%
  filter(n > 0) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill=n)) + theme_fivethirtyeight() + geom_col() +
  xlab(NULL) + coord_flip() + ylab("Word Count") +
  ggtitle("Fear Words Usage in Eisenhower",
          subtitle = "Sentiment Analysis Using NRC")
```



B. BING

1. GET SENTIMENTS

```
#10b) Bing
# Bing categorizes words as Positive or Negative

#10b)(i) Get Sentiments
bing = get_sentiments('bing')
```



	word	sentiment
3188	improperly	negative
3189	impropriety	negative
3190	improve	positive
3191	improved	positive
3192	improvement	positive
3193	improvements	positive
3194	improves	positive
3195	improving	positive
3196	imprudence	negative
3197	imprudent	negative
3198	impudence	negative
3199	impudent	negative
3200	impudently	negative

2. INNER JOIN SENTIMENTS TO TEXT FILE

```
#10b)(ii) Inner Join Sentiments to Text File
tb_un_rm_bing = tb_un_rm %>%
  inner_join(get_sentiments('bing'))
```

topic4.R | Untitled1* | Untitled2* | tidy_df_rm_bing

Filter

	line	word	sentiment
1	1	supreme	positive
2	1	successful	positive
3	2	strong	positive
4	2	denied	negative
5	2	tank	negative
6	2	successful	positive
7	2	uncomfortable	negative
8	2	supreme	positive
9	3	won	positive
10	3	winner	positive
11	3	conflict	negative
12	3	inexpensive	positive
13	3	expensive	negative
14	3	won	positive
15	3	approval	positive
16	3	strong	positive
17	3	support	positive
18	3	supported	positive
19	3	hostile	negative
20	3	crisis	negative
21	3	condemned	negative

Showing 1 to 21 of 34 entries, 3 total columns

All words that have a labeled sentiment are displayed.

3. DO WORD COUNT + INNER JOIN SENTIMENTS TO TEXT FILE

```
#10b)(iii) Do Word Count + Inner Join Sentiments to Text File
tb_un_rm_bing_1 = tb_un_rm_bing %>%
  count(word, sort = TRUE) %>%
  inner_join(tb_un_rm_bing)
```

topic4.R | test.R* | tidy_df_rm_bing1

Filter

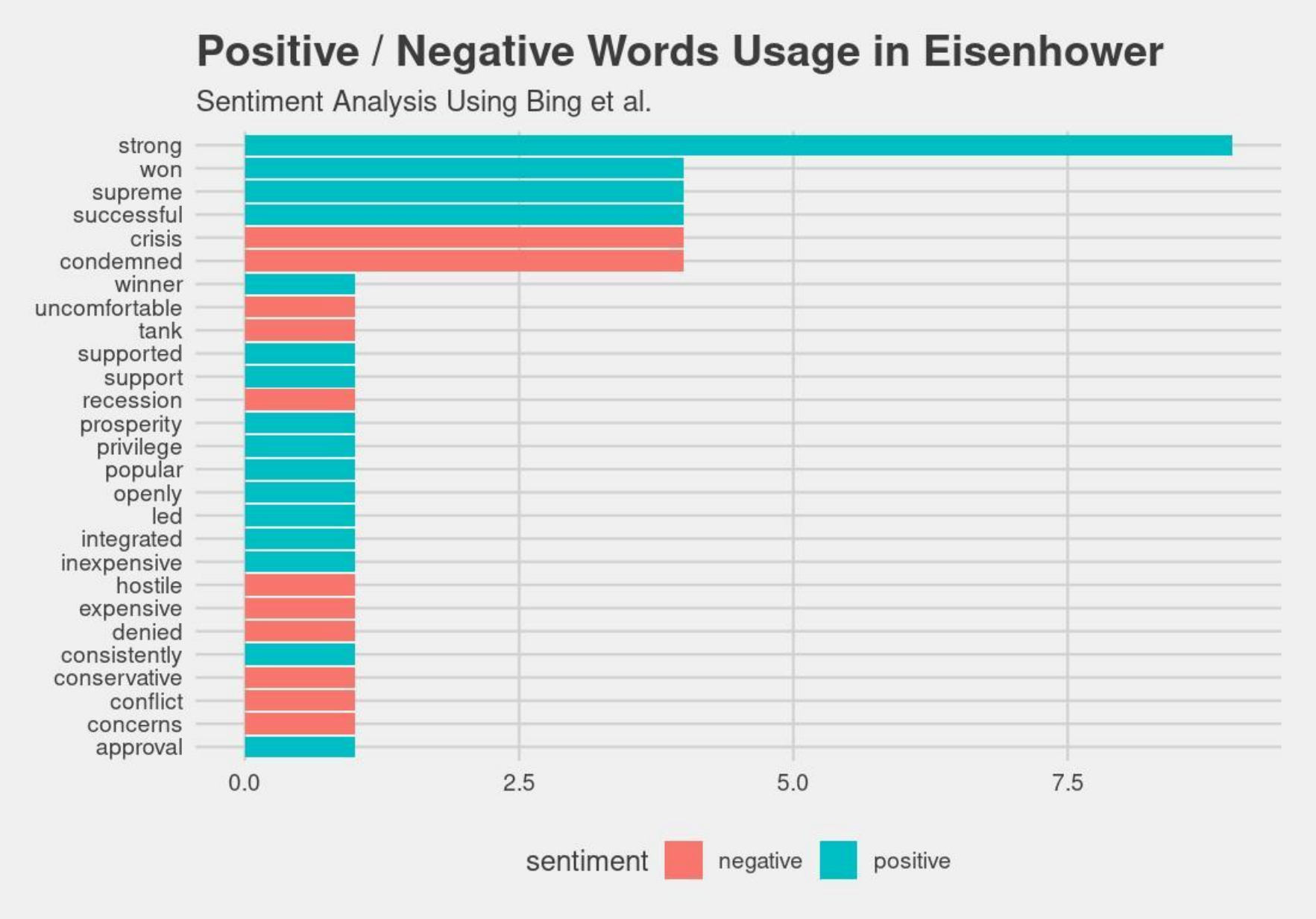
	word	n	line	sentiment
1	strong	3	2	positive
2	strong	3	3	positive
3	strong	3	4	positive
4	condemned	2	3	negative
5	condemned	2	3	negative
6	crisis	2	3	negative
7	crisis	2	3	negative
8	successful	2	1	positive
9	successful	2	2	positive
10	supreme	2	1	positive
11	supreme	2	2	positive
12	won	2	3	positive
13	won	2	3	positive
14	approval	1	3	positive
15	concerns	1	4	negative
16	conflict	1	3	negative
17	conservative	1	4	negative

We created 4 columns (word / n / line / sentiment) by inner joining the "count of words" back to "tb_un_rm_bing".

We need these columns for plotting in the next section.

4. VISUALIZE

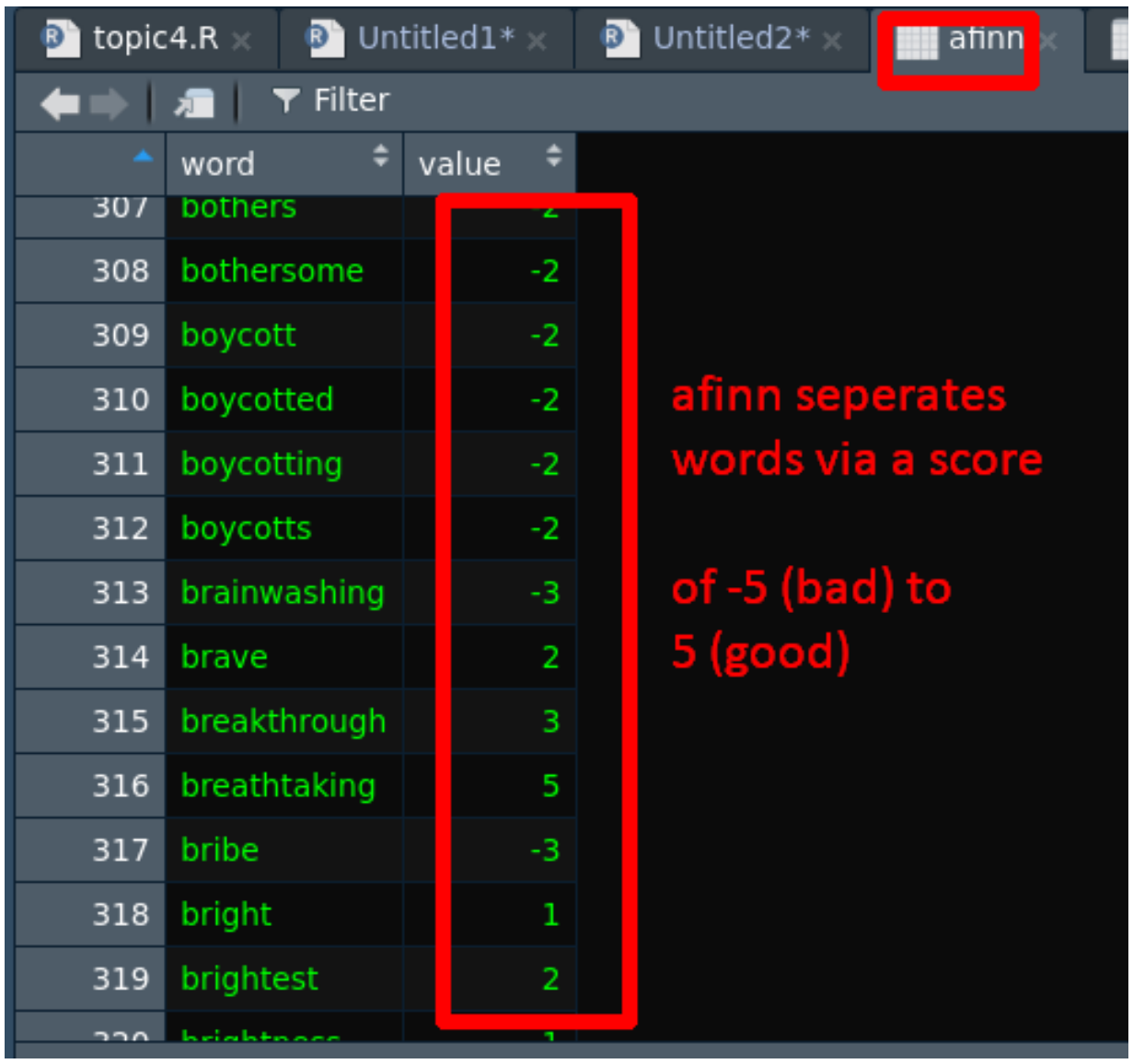
```
#10b)(iv) Visualize
tb_un_rm_bing_1 %>%
  filter(n > 0) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) + theme_fivethirtyeight() + geom_col() +
  xlab(NULL) + coord_flip() + ylab("Word Count") +
  ggtitle("Positive / Negative Words Usage in Eisenhower",
          subtitle = "Sentiment Analysis Using Bing et al.")
```



C. AFINN

1. GET SENTIMENTS

```
#10c) AFINN
# AFINN give word scores
# -5 is Negative while 5 is Positive

#10c)(i) Get Sentiments
afinn = get_sentiments("afinn")
```



	word	value
307	bothers	-2
308	bothersome	-2
309	boycott	-2
310	boycotted	-2
311	boycotting	-2
312	boycotts	-2
313	brainwashing	-3
314	brave	2
315	breakthrough	3
316	breathtaking	5
317	bribe	-3
318	bright	1
319	brightest	2

2. INNER JOIN SENTIMENTS TO TEXT FILE

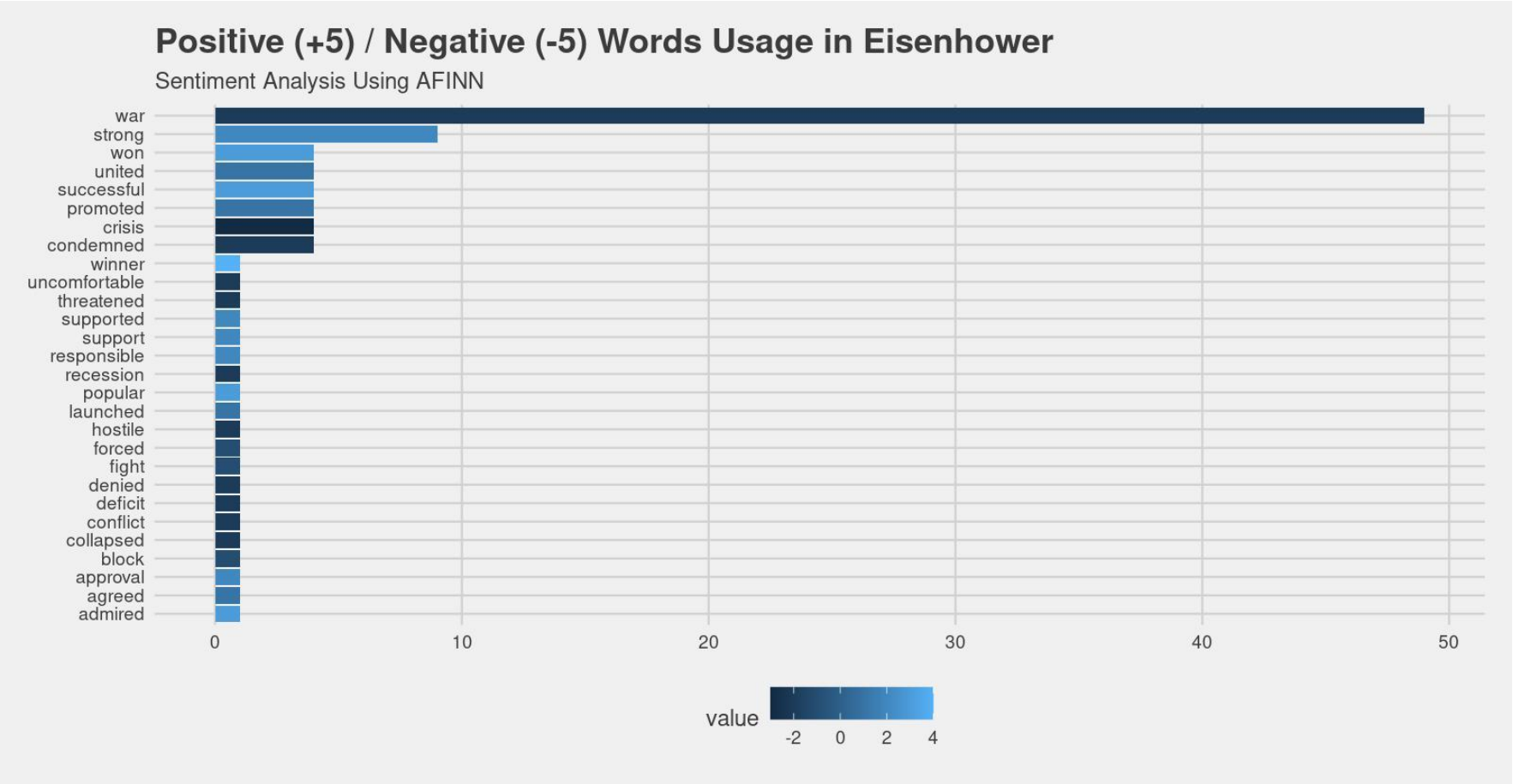
```
#10c)(ii) Inner Join Sentiments to Text File
tb_un_rm_afinn = tb_un_rm %>%
  inner_join(get_sentiments('afinn'))
```

3. DO WORD COUNT + INNER JOIN SENTIMENTS TO TEXT FILE

```
#10c)(iii) Do Word Count + Inner Join Sentiments to Text File
tb_un_rm_afinn_1 = tb_un_rm_afinn %>%
 count(word, sort = TRUE) %>%
 inner_join(tb_un_rm_afinn)
```

4. VISUALIZE

```
#10c)(iv) Visualize
tb_un_rm_afinn_1 %>%
  filter(n > 0) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill=value)) + theme_fivethirtyeight() + geom_col() +
  xlab(NULL) + coord_flip() + ylab("Word Count") +
  ggtitle("Positive (+5) / Negative (-5) Words Usage in Eisenhower",
          subtitle = "Sentiment Analysis Using AFINN")
```



ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.