# EDS 231 - Sentiment_Analysis_II

Juliet Cohen

2022-04-26

Read in Twitter data:

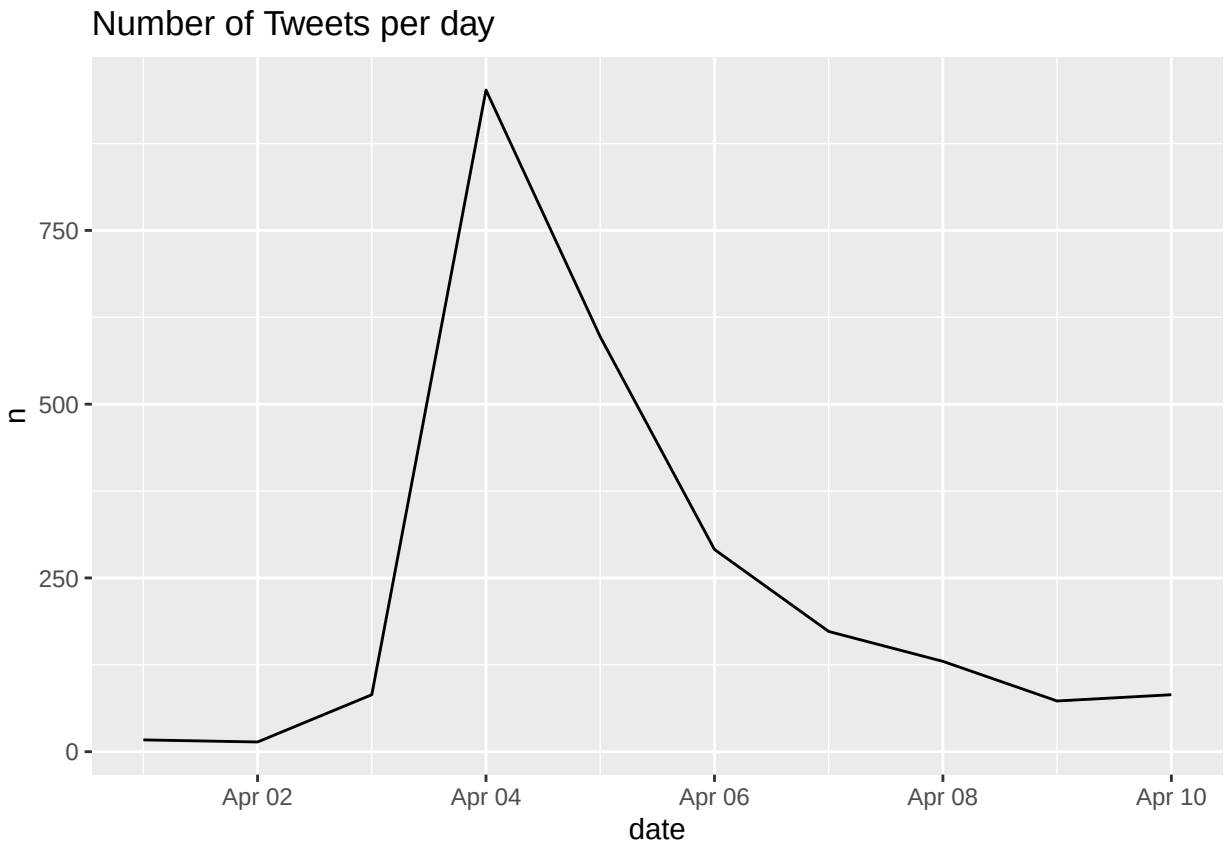
```
raw_tweets <- read.csv("/Users/juliet/Documents/MEDS/Text_Analysis/Text_Analysis/Assignment_3/IPCC_twee

# Extract Date and Title fields
data <- raw_tweets[,c(4,6)]

# create a tibble from the tweet title and date columns
tweets <- tibble(text = data$Title,
                 id = seq(1:length(data$Title)),
                 date = as.Date(data$Date,'%m/%d/%y'))

# head(tweets$text, n = 10)

# simple plot of tweets per day
tweets %>%
  count(date) %>%
  ggplot(aes(x = date, y = n))+
  geom_line() +
  labs(title = "Number of Tweets per day")
```




## Data Cleaning

Think about how to further clean a twitter data set. Let's assume that the mentions of twitter accounts is not useful to us. Remove them from the text field of the tweets tibble.

```
# let's clean up the URLs from the tweets
tweets$text <- gsub("http[^[:space:]]*", "", tweets$text)

# remove emojis
tweets$text <- iconv(tweets$text, "latin1", "ASCII", sub="")

# remove @ and the name of the account tagged because we dont need tagged people in this analysis, just
tweets$text <- gsub("@[^[:space:]]*", "", tweets$text)

# remove # but keep the word following
tweets$text <- gsub("#", "", tweets$text)

# convert all text to lower case
tweets$text <- str_to_lower(tweets$text)

# remove quotes
#tweets$text <- gsub("'", "", tweets$text)

#load sentiment lexicons as usual
bing_sent <- get_sentiments('bing')
nrc_sent <- get_sentiments('nrc')
```

```
# tokenize tweets to individual words so they will be 1 word per row
words <- tweets %>%
  select(id, date, text) %>%
  unnest_tokens(output = word, input = text, token = "words") %>%
  anti_join(stop_words, by = "word") %>% # remove stop words
  left_join(bing_sent, by = "word") %>% # join the words to the sentiment words (label with a sentiment)
  left_join(
    tribble(
      ~sentiment, ~sent_score,
      "positive", 1,
      "negative", -1),
    by = "sentiment")
# the new sentiment score column is numerical, it is how we assign sentiment to words besides just pos/
```

Compare the ten most common terms in the tweets per day. Do you notice anything interesting?

```
# examine trends of the top 10 words per day
top_daily_words <- words %>%
  group_by(date, word) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  slice(1:10)
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.groups`
## argument.
```

```
top_daily_words_table <- kable(top_daily_words,
                          caption = "Top Daily Words by Day")

top_daily_words_table
```

```
# examine trends of the top 10 words total
top_10_total <- top_daily_words %>%
  ungroup() %>%
  slice_max(count, n = 10)

#top_10_total

#order(daily_word_counts$count, decreasing = TRUE)

#subset(daily_word_counts, count == [646)

#unique_word_counts <- unique(sort(daily_word_counts$count, decreasing = TRUE))

#top_daily_words <- head(sort(daily_word_counts$count, decreasing = TRUE), n = 10)
```

Adjust the wordcloud in the "wordcloud" chunk by coloring the positive and negative words so they are identifiable.

Let's say we are interested in the most prominent entities in the Twitter discussion. Which are the top 10 most tagged accounts in the data set. Hint: the "explore_hashtags" chunk is a good starting point.

The Twitter data download comes with a variable called "Sentiment" that must be calculated by Brandwatch. Use your own method to assign each tweet a polarity score (Positive, Negative, Neutral) and compare your classification to Brandwatch's (hint: you'll need to revisit the "raw_tweets" data frame).

Table 1: Top Daily Words by Day

| date | word | count |
|---|---|---|
| 2022-04-01 | ipcc | 13 |
| 2022-04-01 | report | 10 |
| 2022-04-01 | climate | 9 |
| 2022-04-01 | change | 4 |
| 2022-04-01 | carbon | 3 |
| 2022-04-01 | climatereport | 3 |
| 2022-04-01 | fossil | 3 |
| 2022-04-01 | monday | 3 |
| 2022-04-01 | rapid | 3 |
| 2022-04-01 | read | 3 |
| 2022-04-02 | ipcc | 13 |
| 2022-04-02 | report | 11 |
| 2022-04-02 | climate | 7 |
| 2022-04-02 | emissions | 4 |
| 2022-04-02 | 04 | 3 |
| 2022-04-02 | 2022 | 3 |
| 2022-04-02 | carbon | 3 |
| 2022-04-02 | change | 3 |
| 2022-04-02 | gt | 3 |
| 2022-04-02 | monday | 3 |
| 2022-04-03 | ipcc | 107 |
| 2022-04-03 | dr | 75 |
| 2022-04-03 | report | 64 |
| 2022-04-03 | climate | 53 |
| 2022-04-03 | scientists | 30 |
| 2022-04-03 | mitigation | 27 |
| 2022-04-03 | fossil | 26 |
| 2022-04-03 | aitt | 25 |
| 2022-04-03 | authors | 25 |
| 2022-04-03 | dasgupta | 25 |
| 2022-04-04 | ipcc | 646 |
| 2022-04-04 | climate | 634 |
| 2022-04-04 | report | 478 |
| 2022-04-04 | change | 318 |
| 2022-04-04 | world | 170 |
| 2022-04-04 | emissions | 142 |
| 2022-04-04 | scientists | 141 |
| 2022-04-04 | warming | 132 |
| 2022-04-04 | fossil | 105 |
| 2022-04-04 | limit | 101 |
| 2022-04-05 | ipcc | 412 |
| 2022-04-05 | climate | 350 |
| 2022-04-05 | report | 294 |
| 2022-04-05 | change | 157 |
| 2022-04-05 | emissions | 98 |
| 2022-04-05 | world | 74 |
| 2022-04-05 | global | 66 |
| 2022-04-05 | fossil | 61 |
| 2022-04-05 | warming | 59 |
| 2022-04-05 | action | 46 |
| 2022-04-06 | ipcc | 179 |
| 2022-04-06 | climate | 169 |
| 2022-04-06 | report | 121 |
| 2022-04-06 | change | 52 |
| 2022-04-06 | world | 49 |