Topic 3: Sentiment Analysis I

Alex Clippinger

04/17/2022

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

```
library(tidyr)  # text analysis in R
library(lubridate)  # working with date data
library(pdftools)  # read in pdfs
library(tidyverse)
library(tidytext)
library(here)
library(LexisNexisTools)  # Nexis Uni data wrangling
library(sentimentr)
library(readr)
```

Part 0. IPCC Nexis Uni Data Set

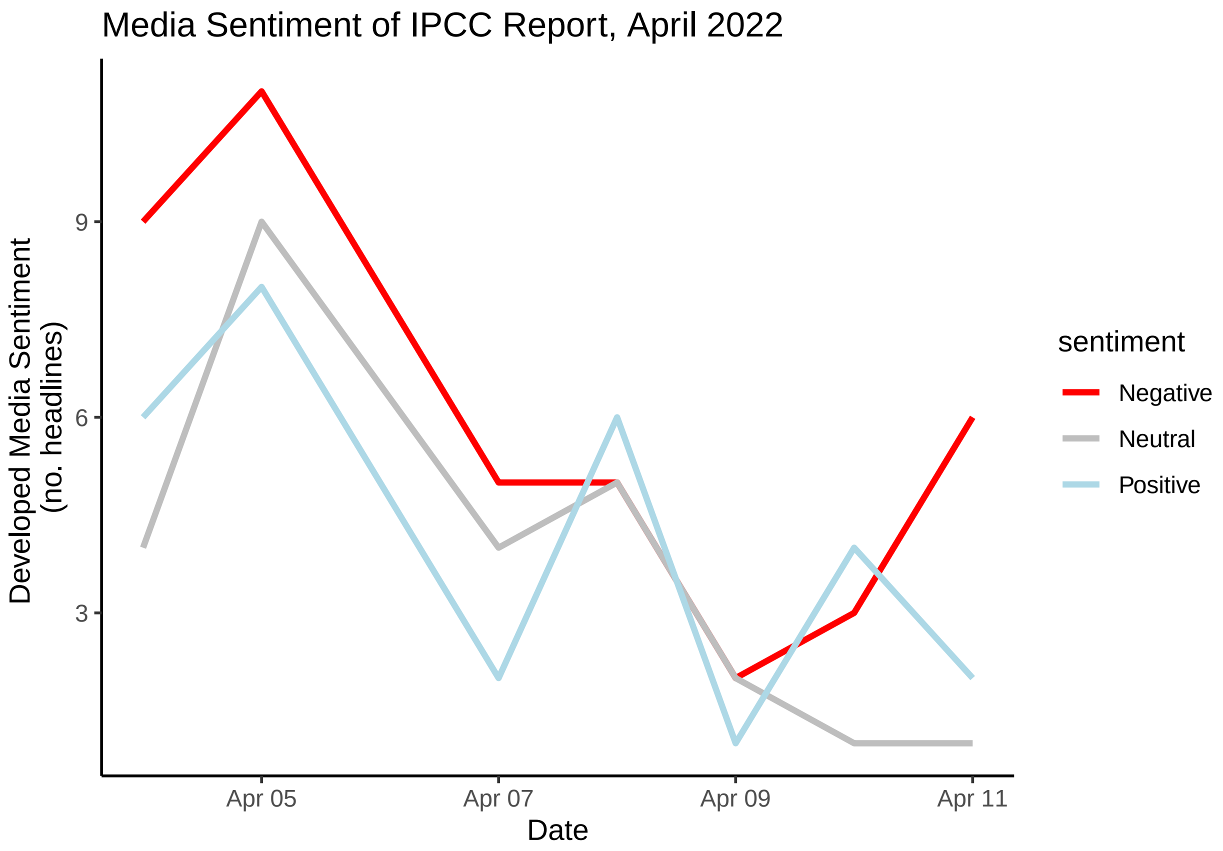
Load Data

```
ipcc_path <- list.files(pattern = "IPCC_results.docx", path = getwd(),
                        full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat <- lnt_read(ipcc_path)  # Object of class 'LNT output'

meta_df <- dat@meta
articles_df <- dat@articles
paragraphs_df <- dat@paragraphs

dat2 <- data_frame(element_id = seq(1:length(meta_df$Headline)),
                   Date = meta_df$Date,
                   Headline = meta_df$Headline)
```

Plot Headline Sentiment

```
mytext <- get_sentences(dat2$Headline)
sent <- sentiment(mytext)

sent_df <- inner_join(dat2, sent, by = "element_id")
```

```
sentiment <- sentiment_by(sent_df$Headline)

sent_df$sentiment <- ifelse(sent_df$sentiment <0, "Negative",
                            ifelse(sent_df$sentiment > 0, "Positive", "Neutral"))

sent_df_plot <- sent_df %>%
  group_by(Date, sentiment) %>%
  summarize(Count = n())

# Create plot similar to Figure 1A. from Froelich et. al.
ggplot(sent_df_plot, aes(x = Date, y = Count, color = sentiment)) +
  geom_line(size = 1.1) +
  scale_color_manual(values = c("Red", "Grey", "lightblue")) +
  theme_classic() +
  labs(y = "Developed Media Sentiment \n(no. headlines)",
       title = "Media Sentiment of IPCC Report, April 2022")
```



Part 1-3. Download Search Results From Nexis Uni Database

For this assignment, I chose to search for “Ketanji Brown Jackson”, who was recently nominated to serve in the U.S. Supreme Court. I chose this search term because I expected mixed sentiment after the nomination hearings turned into, as one article put it, a “brawl”.

Part 4. Read the Nexis article document into R

```r
kbj_path <- list.files(pattern = "KBJ_results.docx", path = getwd(),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

kbj_dat <- lnt_read(kbj_path)  # Object of class 'LNT output'

kbj_meta_df <- kbj_dat@meta
kbj_articles_df <- kbj_dat@articles
kbj_paragraphs_df <- kbj_dat@paragraphs

kbj_dat2 <- data_frame(element_id = seq(1:length(kbj_meta_df$Headline)),
                       Date = kbj_meta_df$Date,
                       Headline = kbj_meta_df$Headline)

# Using full text from the articles
kbj_paragraphs_dat <- data_frame(element_id = kbj_paragraphs_df$Art_ID,
                             Text = kbj_paragraphs_df$Paragraph)

kbj_dat3 <- inner_join(kbj_dat2, kbj_paragraphs_dat, by = "element_id")
```

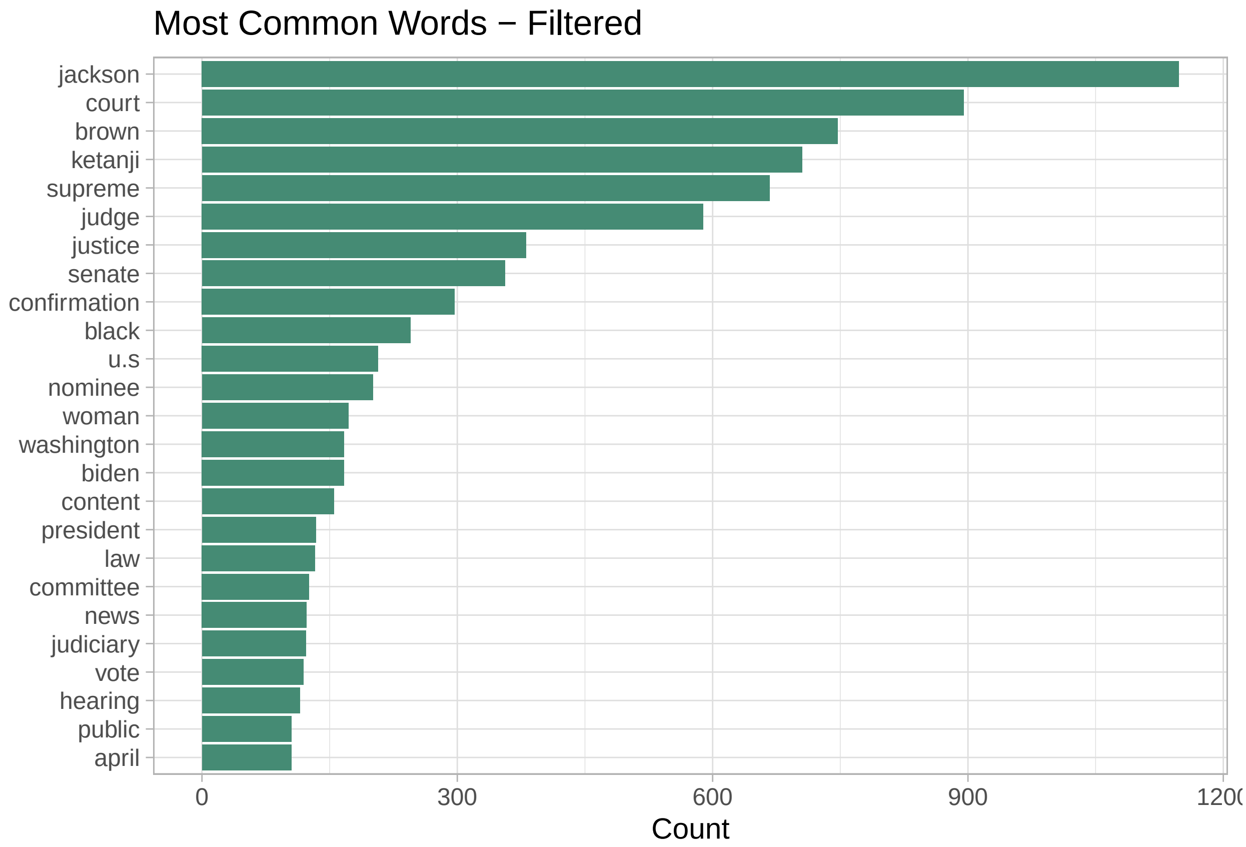
Part 5. Clean Nexis Data

```r
bing_sent <- get_sentiments('bing')  # grab bing sentiment lexicon from tidytext

text_words <- kbj_dat3 %>%
  unnest_tokens(output = word, input = Text, token = "words")
```

```r
# Remove stop words
sent_words <- text_words %>%
  anti_join(stop_words, by = "word")

# Clean remaining data
clean_tokens <- str_remove_all(sent_words$word, "[:digit:]") # remove numbers
clean_tokens <- str_remove_all(clean_tokens, '\\([^()]*\\)') # remove parentheses
clean_tokens <- str_remove_all(clean_tokens, 'http\\S+') # remove urls
clean_tokens <- gsub("'s", '', clean_tokens) # remove possessive

sent_words$word <- clean_tokens  # remove all numbers

# Remove empty strings and added stop words
tib <- subset(sent_words, !(word %in% c("", ".", "mstruck", "newstex")))

# Reassign
sent_words <- tib
```

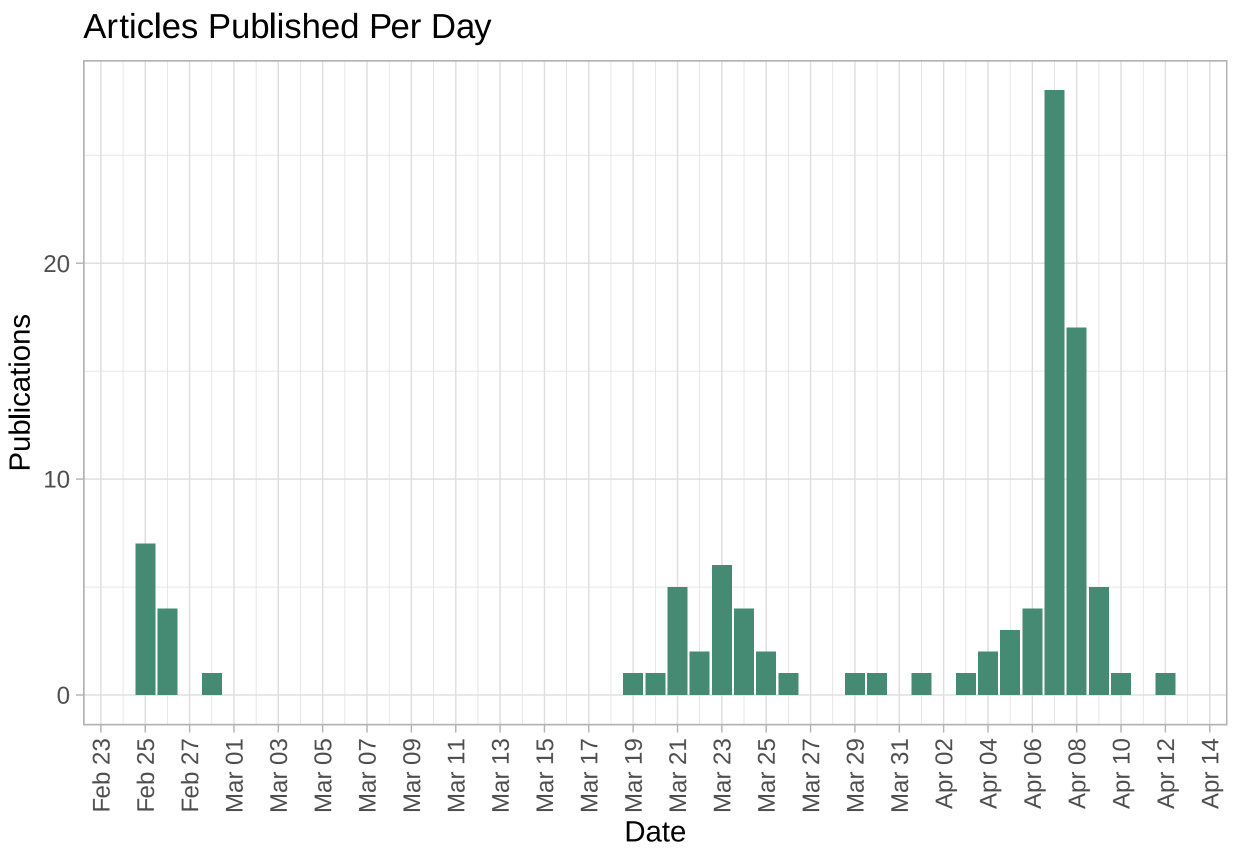
Part 6. Data Exploration

```r
sent_words_plot <- sent_words %>%
  count(word, sort = TRUE) %>%
  mutate(word = reorder(word, n)) %>%
  head(25)

ggplot(data = sent_words_plot, aes(n, word)) +
  geom_col(fill="aquamarine4") +
  labs(y = NULL, x = "Count", title="Most Common Words - Filtered") +
  theme_light()
```



The top 25 words above seem intuitive.

```r
pub_plot <- sent_words %>%
  group_by(Date) %>%
  summarize(Publications = n_distinct(element_id))

ggplot(pub_plot, aes(Date, Publications)) +
  geom_bar(fill="aquamarine4", stat = "identity") +
  labs(title="Articles Published Per Day") +
  scale_x_date(date_labels = "%b %d", date_breaks = "2 days") +
  guides(x = guide_axis(angle = 90)) +
  theme_light()
```



Most articles were published on April 7th, when Ketanji Brown Jackson was confirmed to the Supreme Court. One notable trend in the plot above is the break in coverage between February 28th to March 19th, representing the time between the nomination was announced to the beginning of the hearings.

Part 7. Plot Emotion Words

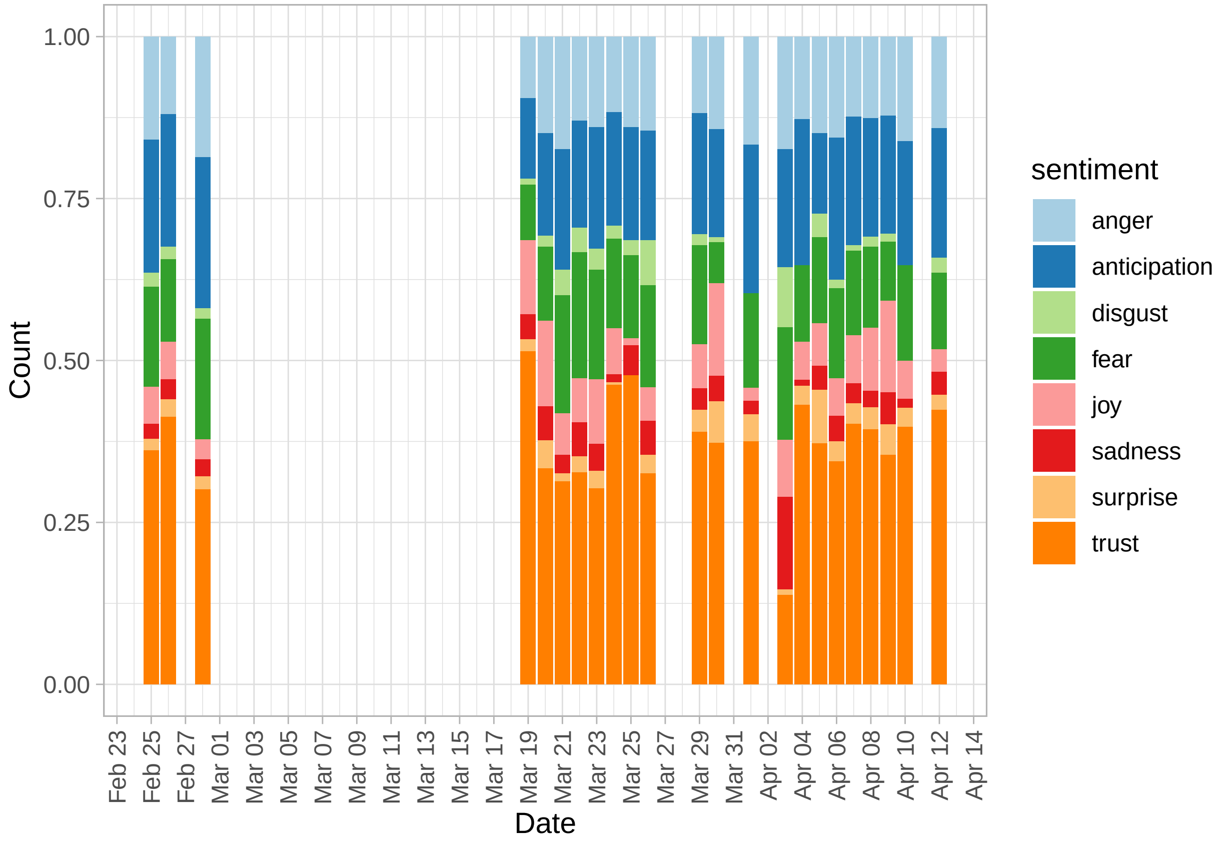
Get Sentiment Words

```
nrc_word_counts <- sent_words %>%
  select(-Headline) %>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!(sentiment %in% c("positive", "negative")))
```

Plot Results

```
nrc_words_plot <- nrc_word_counts %>%
  filter(!is.na(Date)) %>%
  group_by(Date, sentiment) %>%
  summarize(Count = n()) %>%
  ungroup()
```

```
day_sum <- nrc_words_plot %>%
  group_by(Date) %>%
  summarize(Day_Sum = sum(Count))

nrc_words_plot <- nrc_words_plot %>%
  left_join(day_sum, by = "Date") %>%
  mutate(Emotion_Pct = Count/Day_Sum) %>%
  arrange(Date, Count)

ggplot(data = nrc_words_plot, aes(fill = sentiment, y = Count, x = Date)) +
  geom_bar(position = "fill", stat = "identity") +
  scale_fill_brewer(palette = "Paired") +
  scale_x_date(date_labels = "%b %d", date_breaks = "2 days") +
  guides(x = guide_axis(angle = 90)) +
  theme_light()
```



Overall, the sentiment remained relatively the same, with the most common sentiment per day being “trust”, followed by “anticipation”. However, the contentious nature of the nomination hearings is reflected in the sentiment of the days preceding April 7th, when Ketanji Brown Jackson was confirmed to the Supreme Court. Notably, on April 3rd, the sentiments “sadness” and “fear” saw an increased percentage. As we saw during data exploration, there was only one article in our data set published on that date, titled: “Opinion: Ketanji Brown Jackson is a disgrace to American justice. The Griffon News: Missouri Western State College.” Given this was an opinion piece that clearly opposed the judge, it makes sense that the distribution of sentiment looks markedly different than other days.