EDS 231 Week 3 Lab

Wylie Hampson

4/19/2022

Using the "IPCC" Nexis Uni data set from the class presentation and the pseudo code we discussed, recreate Figure 1A from Froelich et al. (Date x # of 1) positive, 2) negative, 3) neutral headlines):

```
setwd(here())

my_files <- list.files(pattern = ".docx", path = here("data", "IPCC_nexis"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat <- lnt_read(my_files)


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

# Use the words from the article headlines
dat2<- data_frame(element_id = seq(1:length(meta_df$Headline)),
                  Date = meta_df$Date,
                  Headline = meta_df$Headline)
```

Get the IPCC files from the docx into R

```
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 %>%
  arrange(sentiment)
```

Get sentiment values for each headline.

```
## # A tibble: 109 x 6
```

```
##    element_id Date       Headline               sentence_id word_count sentiment
##         <int> <date>     <chr>                        <int>      <int>     <dbl>
##  1         66 2022-04-04 Scientists risk arres~           1          7    -0.756
##  2         91 2022-04-07 The 'climate change' ~           1          9    -0.75
##  3         28 2022-04-09 The Dread 1.5 Degree ~           1          6    -0.714
##  4         43 2022-04-06 India's banks unprepa~           1          7    -0.510
##  5         34 2022-04-08 Dangerous radicals ar~           1          6    -0.449
##  6         14 2022-04-04 'Now or never' to avo~           1          8    -0.442
##  7         78 2022-04-07 Statewide Gas Ban Bil~           1         10    -0.427
##  8         50 2022-04-04 Guardian: Media 'Bare~           1          8    -0.407
##  9         62 2022-04-06 Governor Youngkin's I~           1         11    -0.377
## 10          7 2022-04-05 Narrow path to avoid ~           1          8    -0.354
## # ... with 99 more rows
```

```
sent_df$polarity <- ifelse(sent_df$sentiment <0, -1,
                           ifelse(sent_df$sentiment > 0, 1, 0))
```

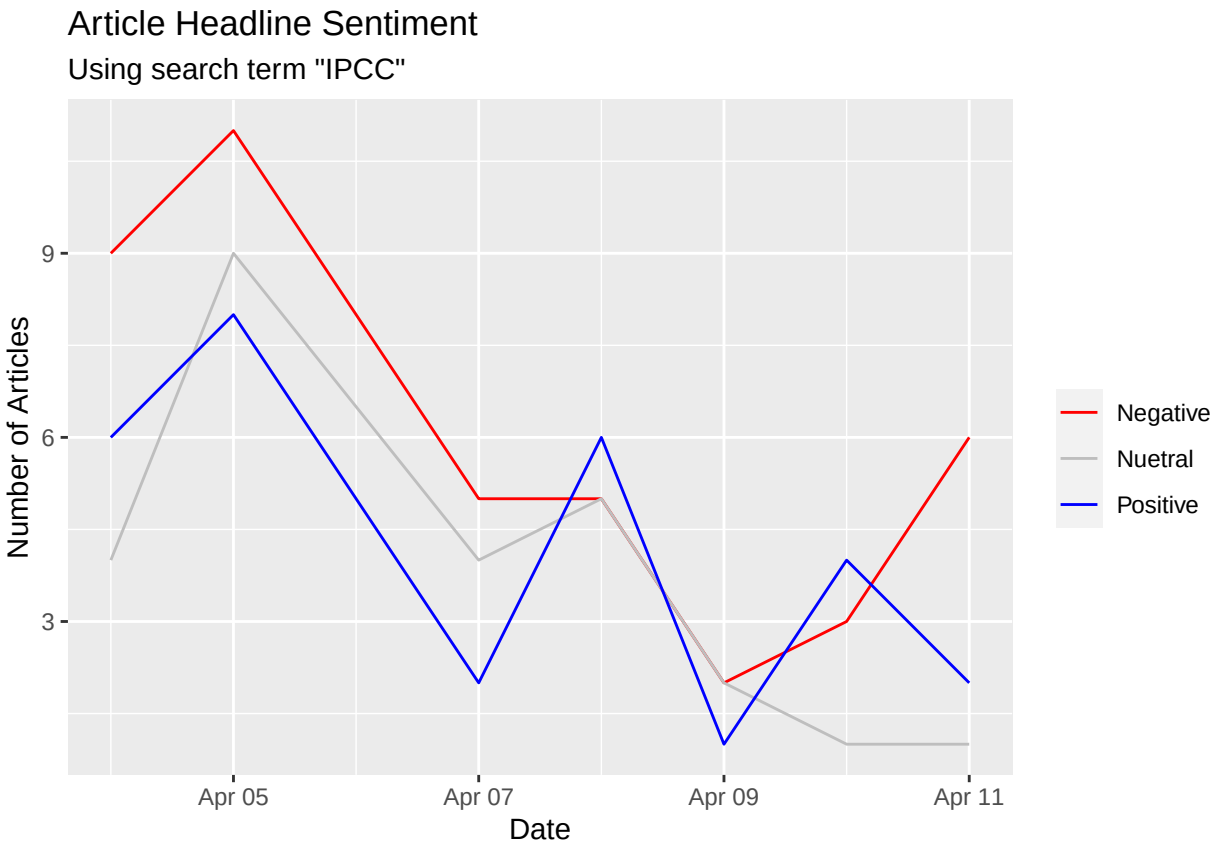
Assign polarity to positive (1), negative(-1), or neutral (0) by rounding up if > 0 and down if < 0.

```
plot_df <- sent_df %>%
  group_by(Date, polarity) %>%
  summarise(count = n())
```

Plot the count of the polarity.

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

```
ggplot(data = plot_df, aes(x = Date, y = count)) +
  geom_line(aes(col = as.factor(polarity))) +
  scale_colour_manual(labels = c("Negative","Nuetral", "Positive"),
                      values = c("red", "gray", "blue")) +
  labs(color = "",
       title = "Article Headline Sentiment",
       subtitle = "Using search term \"IPCC\"",
       x = "Date",
       y = "Number of Articles")
```



Now access the Nexis database and use search term "wind power" to pull articles.

I pulled 100 articles from the search term "wind power" between dates 01/29/2022 and 04/19/2022. Here I'll look at the words from the whole articles instead of just the headlines.

```
my_files_wind <- list.files(pattern = ".docx", path = here("data", "wind_power_nexis"),
                        full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat_wind <- lnt_read(my_files_wind)

meta_df_wind <- dat_wind@meta
articles_df_wind <- dat_wind@articles
paragraphs_df_wind <- dat_wind@paragraphs

dat2_wind<- data_frame(element_id = seq(1:length(meta_df_wind$Headline)),
                       Date = meta_df_wind$Date,
                       Headline = meta_df_wind$Headline)

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

dat3_wind <- inner_join(dat2_wind, paragraphs_dat_wind, by = "element_id")
```

```
nrc_sentiment <- get_sentiments("nrc")
```

Pull in the nrc sentiment dataframe

```
text_words <- dat3_wind %>%
  unnest_tokens(output = word, input = Text, token = "words")

word_sentiments <- text_words %>%
  anti_join(stop_words, by = "word") %>%
  inner_join(nrc_sentiment, by = "word") %>%
  filter(sentiment != "positive" & sentiment != "negative")
```

Unnest the article words and join the nrc sentiments to the dataframe.

```
sentiment_count <- word_sentiments %>%
  group_by(Date, sentiment) %>%
  summarise(count = n())
```

Count how many words there are for each sentiment for each day.

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

```
daily_sentiments <- sentiment_count %>%
  select(Date, count) %>%
  group_by(Date) %>%
  summarise(total_sentiments = sum(count))
```

Add a column to get the total sentiment words for that day so we can find a percent.

```
plot_df_wind <- left_join(sentiment_count, daily_sentiments) %>%
  mutate(sentiment_percent = count / total_sentiments * 100)
```

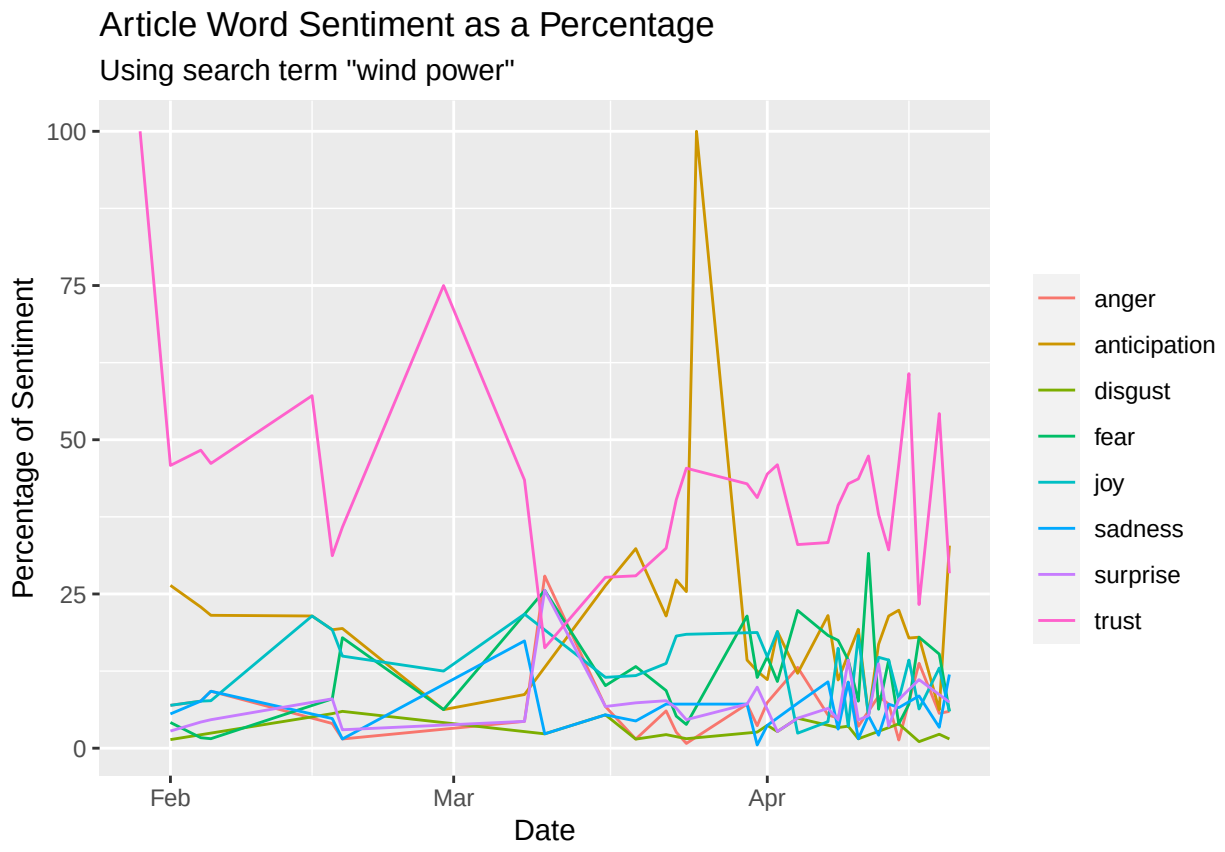
Find the percentage of each sentiment words for each day.

```
## Joining, by = "Date"
```

```
ggplot(data = plot_df_wind, aes(x = Date, y = sentiment_percent)) +
  geom_line(aes(col = as.factor(sentiment))) +
  labs(color = "",
       title = "Article Word Sentiment as a Percentage",
       subtitle = "Using search term \"wind power\"",
       x = "Date",
       y = "Percentage of Sentiment")
```

Plot the sentiment percentages by the date.

```
## Warning: Removed 8 row(s) containing missing values (geom_path).
```



How does the distribution of emotion words change over time? Can you think of any reason this would be the case?

It was really interesting to look at the sentiment of words from articles related to "wind power". We can see the the sentiment that is generally most common throughout the whole period is trust. This makes sense to me because generally I think most people trust wind power and support it. However it still does flucuate and get higher and lower at times. I also think that becasue these are articles, trust might be really high because authors want to convince people to trust wind power and get on board with it. Another sentiment that seems pretty common is anticipation. I think this makes sense because I think people really want wind power to become more common. There is an interesting spike for anticipation on March 25th, leading you to believe that something got people really excited about wind power. However, what it looks like really happened is that there were very few articles that day, and the only words with sentiment had to do with anticipation, so it's a bit misleading.