Lab 2

Felicia Cruz

4/19/2022

0. Recreating Figure 1A from Froelich et al.

- group_by date
- summarize mean sentiment for that day
- plot x = day, y = sentiment

```
# load data
ipcc_files <- list.files(pattern = ".docx", path = here("labs", "data", "ipcc_nexis"),
                         full.names = TRUE, recursive = TRUE, ignore.case = TRUE)


ipcc_dat <- lnt_read(ipcc_files)


ipcc_meta_df <- ipcc_dat@meta
ippc_articles_df <- ipcc_dat@articles
ippc_paragraphs_df <- ipcc_dat@paragraphs

# headline df
ipcc_dat_2<- data_frame(element_id = seq(1:length(ipcc_meta_df$Headline)), Date = ipcc_meta_df$Date, Hea
```

```
mytext <- get_sentences(ipcc_dat_2$Headline)
sent <- sentiment(mytext)

sent_df <- inner_join(ipcc_dat_2, sent, by = "element_id")

sentiment <- sentiment_by(sent_df$Headline)

sent_df %>%
  arrange(sentiment)
```

```
## # 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
```

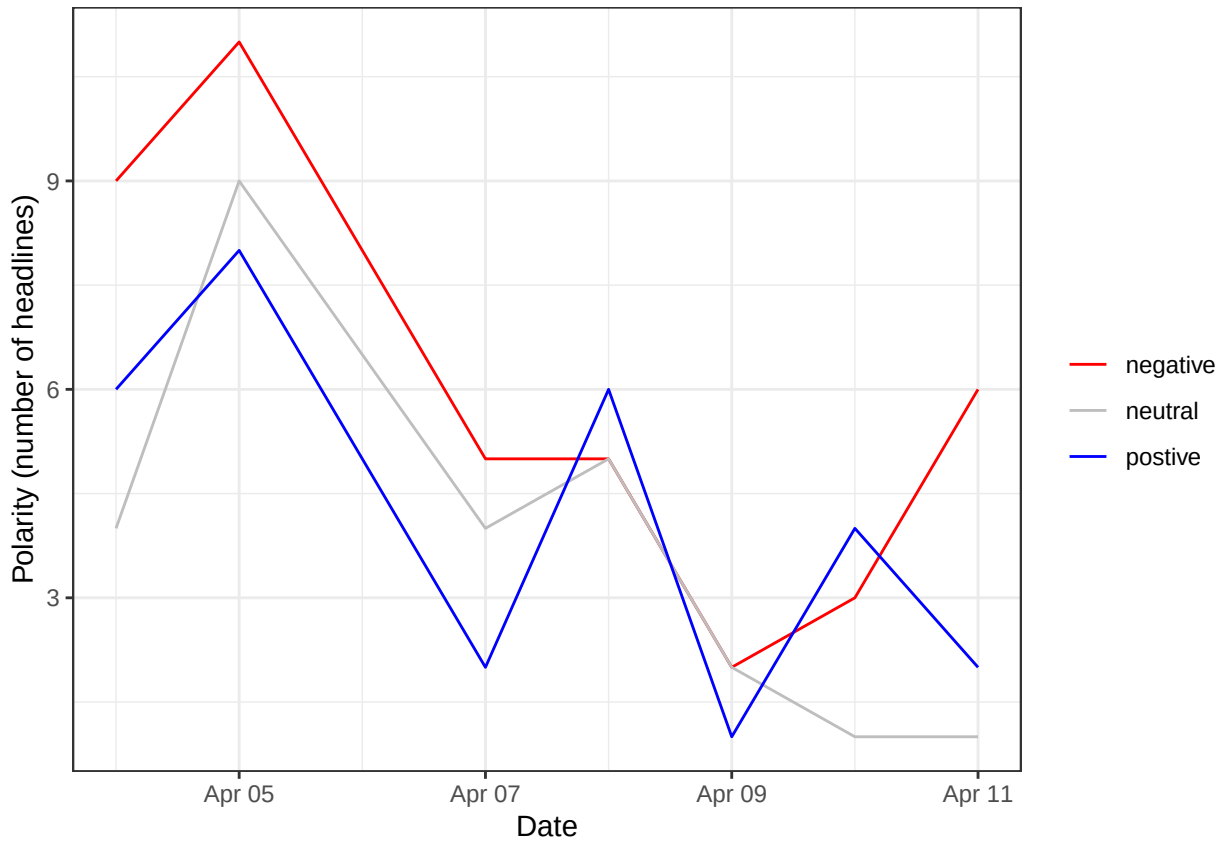
```
##  7         78 2022-04-07 Statewide Gas Ban Bil~           1         10    -0.427
##  8         50 2022-04-04 Guardian: Media 'Bare~           1          8    -0.407
##  9         62 2022-04-06 Governor Youngkin's I~           1         11    -0.377
## 10          7 2022-04-05 Narrow path to avoid ~           1          8    -0.354
## # ... with 99 more rows
```

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

```
sent_df_summary <- sent_df %>%
  group_by(Date, polarity) %>%
  summarize(count_polarity_date = n())
```

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

```
ggplot(data = sent_df_summary, aes(x = Date, y = count_polarity_date, )) +
  geom_line(aes(group = as.factor(polarity), color = as.factor(polarity))) +
  scale_color_manual(labels = c("negative", "neutral", "postive"), values = c("red", "gray", "blue")) +
  labs(color = "",
       y = "Polarity (number of headlines)") +
  theme_bw()
```



1. Nexis Uni database

For my Nexis Uni database query I chose the search term "wildfire" with the date range April 1, 2021 to July 1, 2021 to encapsulate articles both leading up to and at the start of wildfire season. I pulled 100 full

text results for this lab and the final plots span April 15, 2021 to July 1, 2021.

```
my_files <- list.files(pattern = ".DOCX", path = here("labs", "data", "wildfire_2_nexis"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

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

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

```
# use the full text from the articles
dat2<- data_frame(element_id = seq(1:length(meta_df$Headline)), Date = meta_df$Date, Headline = meta_df$

paragraphs_dat <- data_frame(element_id = paragraphs_df$Art_ID, Text  = paragraphs_df$Paragraph)


dat3 <- inner_join(dat2, paragraphs_dat, by = "element_id")
```

```
nrc_sent <- get_sentiments('nrc') #requires downloading a large dataset via prompt
```

```
text_words <- dat3  %>%
  unnest_tokens(output = word, input = Text, token = 'words') # take the paragraphs and unnest to get a

sent_words <- text_words %>%
  anti_join(stop_words, by = 'word') %>% # remove stop words
  inner_join(nrc_sent, by = 'word') # join and retain only sentiment words
```

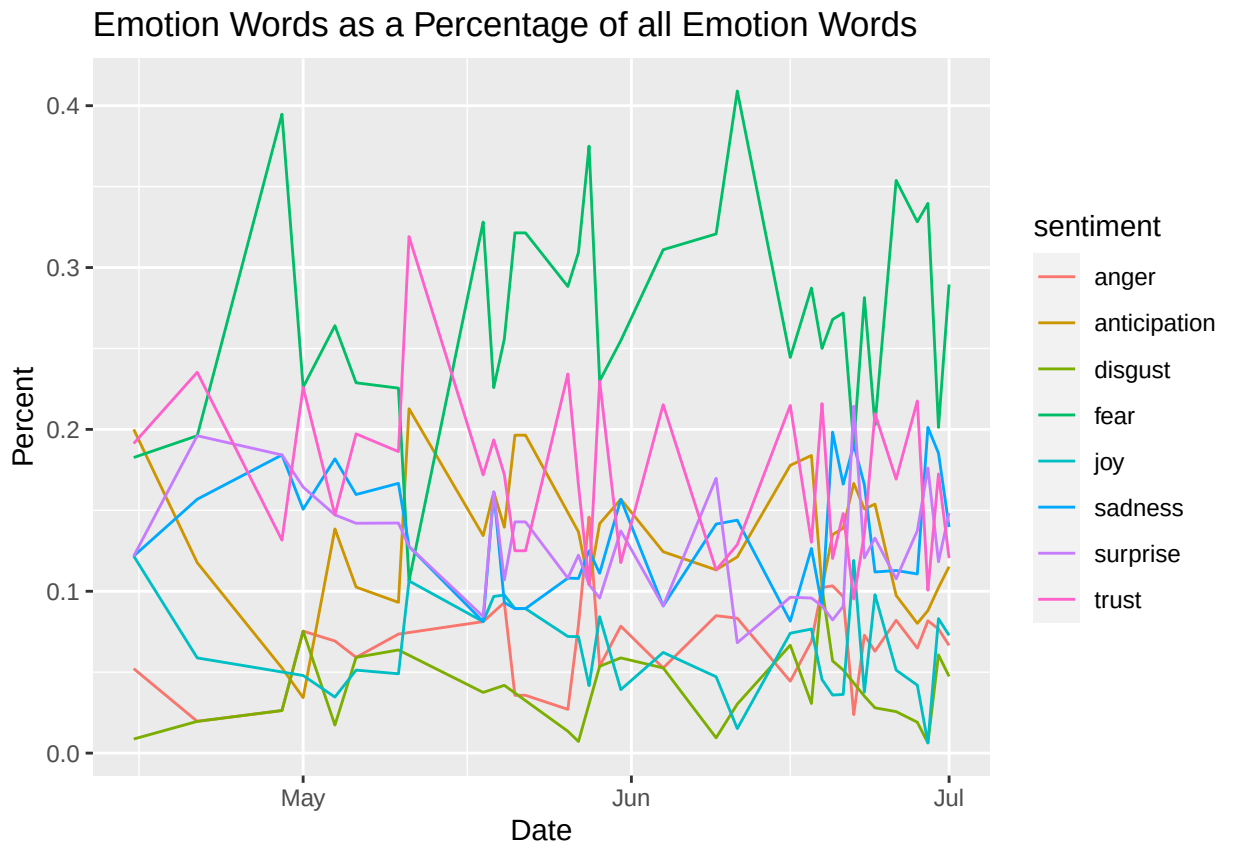
```
sentiment_counts <- sent_words %>%
  filter(sentiment != "positive",
         sentiment != "negative") %>%
  group_by(Date, sentiment) %>%
  summarize(count_by_sentiment = n())
```

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

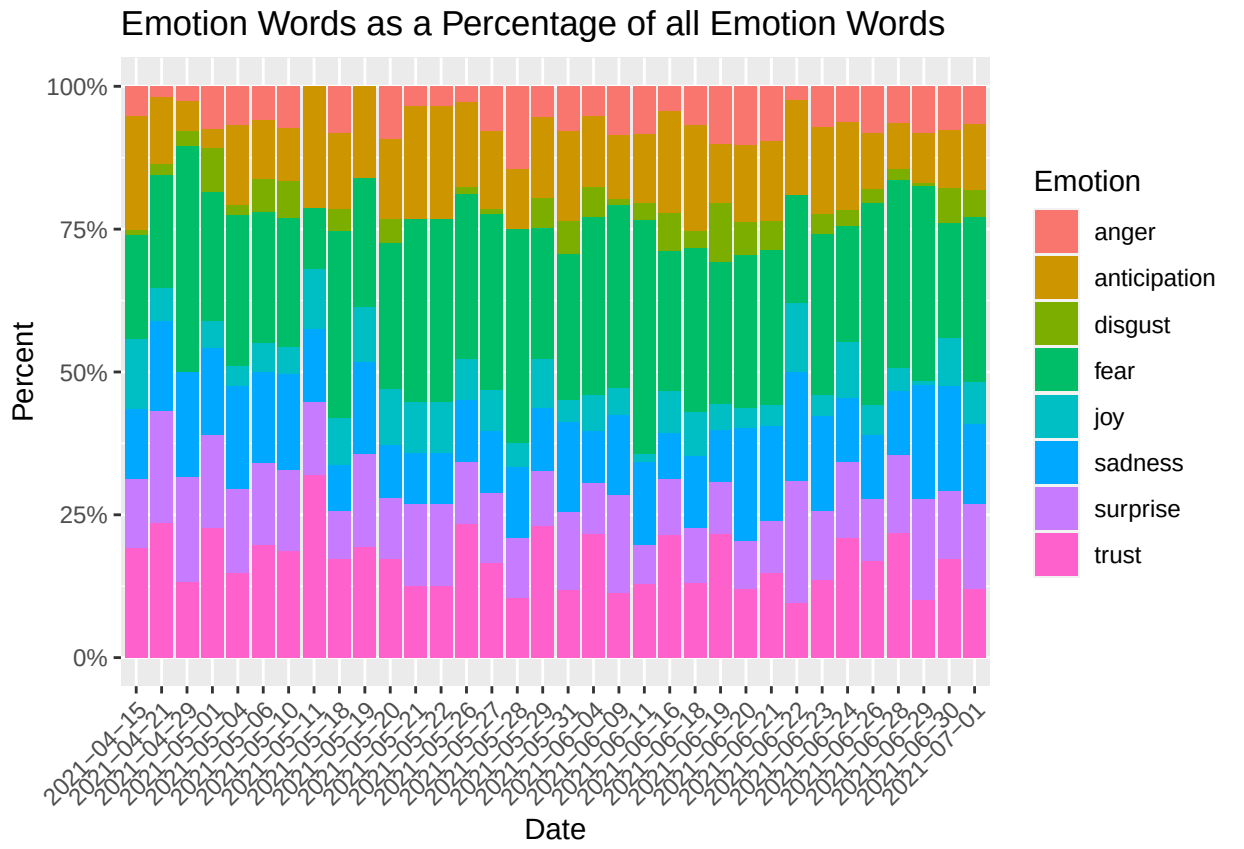
```
daily_sent_words_df <- sentiment_counts %>%
 # select(-sentiment) %>%
  group_by(Date) %>%
  summarize(total_daily_sent = sum(count_by_sentiment))

# join the two dfs by date
sent_words_summary <- left_join(sentiment_counts, daily_sent_words_df, by = "Date")

# add a percentage column
sent_words_summary <- sent_words_summary %>%
  mutate(percent = count_by_sentiment / total_daily_sent) %>%
  filter(!is.na(Date))
```

Plot the amount of emotion words (the 8 from nrc) as a percentage of all the emotion words used each day (aggregate text from articles published on the same day).

```
# line graph with 8 lines
ggplot(data = sent_words_summary, aes(x = Date, y = percent)) +
  geom_line(aes(group = sentiment, color = sentiment)) +
  labs(x = "Date",
       y = "Percent",
       fill = "Emotion",
       title = "Emotion Words as a Percentage of all Emotion Words")
```




```
# stacked bar chart
ggplot(data = sent_words_summary, aes(x = as.factor(Date), y = percent, fill = sentiment)) +
  geom_col(position = "stack") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Date",
       y = "Percent",
       fill = "Emotion",
       title = "Emotion Words as a Percentage of all Emotion Words") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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

From the stacked bar chart, we can see that fear is often the highest percentage of emotion words on a given day. While this is often the case, from the line graph we see that there are big peaks and dips in the line for fear. Anticipation, while normally making up a smaller percentage of emotion words compared to fear, also shows many peaks and dips, often aligning with those of fear. This is somewhat intuitive or expected since the anticipation of wildfires can go hand-in-hand with fear. All of these fluctuations could be reflecting articles written in the days leading up to or surrounding a wildfire event. While this time frame covers some of the wildfire season, the number of wildfire events and intensities will vary over the season and this is one reason that I think would contribute to the distribution of emotion word percentages.