Lab_2

Allie Cole

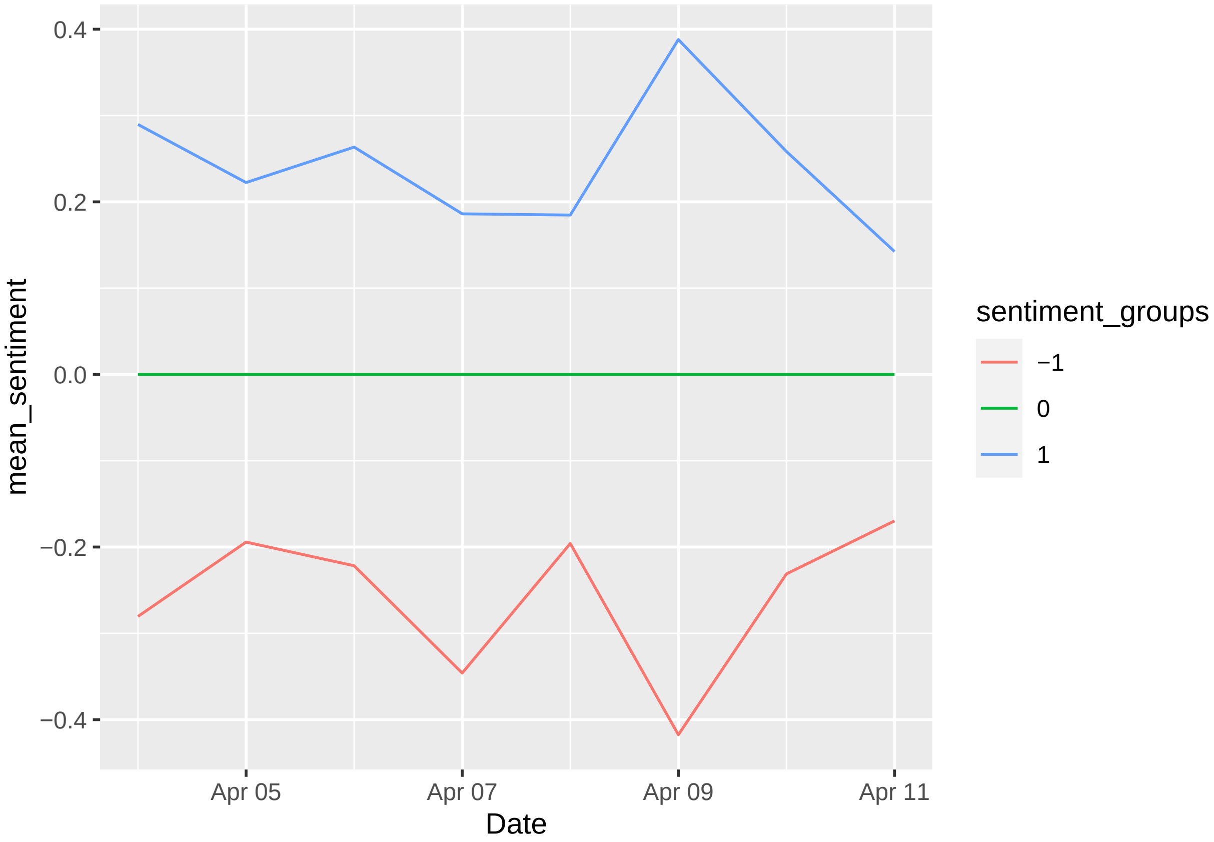
4/18/2022

This is the graph from the end of class

```
sent_df %>%
  mutate(sentiment_groups = case_when(sentiment > 0 ~ "1",
                                      sentiment == 0 ~ "0",
                                      sentiment < 0 ~ "-1"),
         factor(sentiment_groups, levels = c(1, 0, -1))) %>%
  group_by(Date, sentiment_groups) %>%
  summarise(mean_sentiment = mean(sentiment)) %>%
  ggplot(aes(x = Date,
             y = mean_sentiment,
             color = sentiment_groups)) +
  geom_line(position = "dodge")
```

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

```
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```



#Now we load in Marine Ecology Data and finish the homework

```
my_files <- list.files(pattern = ".docx", path = here("Lab_2", "data"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)
dat <- lnt_read(my_files) #Object of class 'LNT output'
```

```
## Warning in lnt_asDate(date.v, ...): More than one language was detected. The
## most likely one was chosen (English 98%)
```

```
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)

#May be of use for assignment: using the full text from the articles
paragraphs_dat <- data_frame(element_id = paragraphs_df$Art_ID, Text  = paragraphs_df$Paragraph)

dat3 <- inner_join(dat2, paragraphs_dat, by = "element_id")
#this is get rid of any links and anything that contains less than 20 words
cleanpars <- dat3 %>%
  mutate(link = str_detect(dat3$Text, "http", negate = TRUE)) %>%
  filter(link == TRUE & nchar(dat3$Text) > 20)
```

```r
mytext <- get_sentences(cleanpars$Text)
sent <- sentiment(mytext)
```

```r
sent_df <- inner_join(cleanpars, sent, by = "element_id")
sentiment <- sentiment_by(sent_df$Text)
```

```
## Warning: Each time `sentiment_by` is run it has to do sentence boundary disambiguation when a
## raw `character` vector is passed to `text.var`. This may be costly of time and
## memory.  It is highly recommended that the user first runs the raw `character`
## vector through the `get_sentences` function.
```

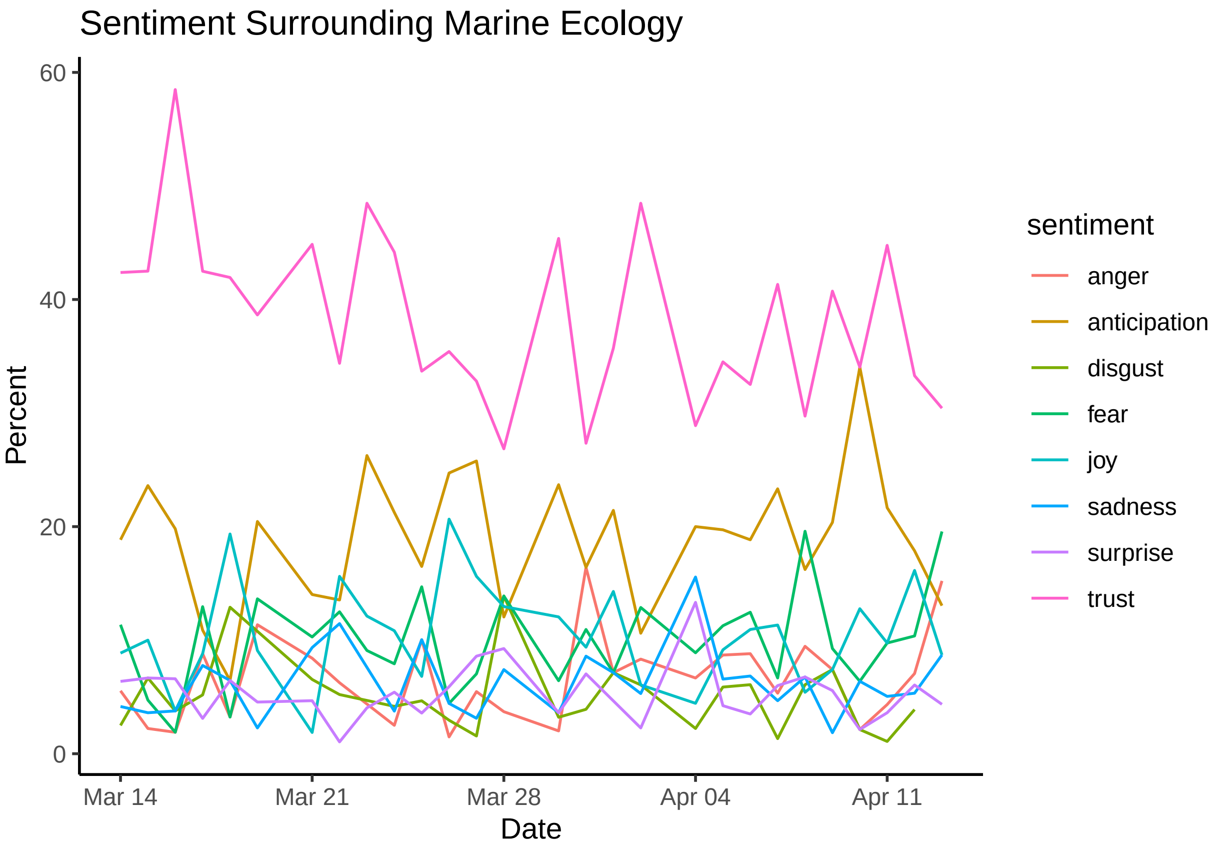
```r
sent_df %>%
  arrange(sentiment)
```

```
## # A tibble: 2,001 x 8
##    element_id Date       Headline  Text   link  sentence_id word_count sentiment
##         <int> <date>     <chr>     <chr>  <lgl>       <int>      <int>     <dbl>
##  1         79 2022-04-05 Nature C~ "(TNS~ TRUE            2         28    -0.378
##  2         79 2022-04-05 Nature C~ "The ~ TRUE            2         28    -0.378
##  3         79 2022-04-05 Nature C~ "The ~ TRUE            2         28    -0.378
##  4         79 2022-04-05 Nature C~ "\"Ke~ TRUE            2         28    -0.378
##  5         79 2022-04-05 Nature C~ "Prio~ TRUE            2         28    -0.378
##  6         79 2022-04-05 Nature C~ "\"Ke~ TRUE            2         28    -0.378
##  7         79 2022-04-05 Nature C~ "The ~ TRUE            2         28    -0.378
##  8         79 2022-04-05 Nature C~ "\"We~ TRUE            2         28    -0.378
##  9         79 2022-04-05 Nature C~ "Drs.~ TRUE            2         28    -0.378
## 10         79 2022-04-05 Nature C~ "\"Th~ TRUE            2         28    -0.378
## # ... with 1,991 more rows
```

```r
nrc_sent <- get_sentiments('nrc')

text_words <- cleanpars  %>%
  unnest_tokens(output = word, input = Text, token = 'words')

sent_words <- text_words %>% #break text into individual words
  anti_join(stop_words, by = 'word') %>%
  inner_join(nrc_sent, by = 'word') %>%
  filter(!sentiment %in% c("positive", "negative")) %>%
  mutate(Date = as_date(Date))

sent_word_count <- sent_words %>%
  group_by(Date, sentiment) %>%
  count(sentiment) %>%
  ungroup() %>%
  group_by(Date) %>%
  mutate(n_max = sum(n),
         percent = round((n / n_max) * 100, 2))
```

```r
ggplot(data = sent_word_count) +
  geom_line(aes(x = Date, y = percent, color = sentiment)) +
  theme_classic() +
  labs(title = "Sentiment Surrounding Marine Ecology",
       y = "Percent",
       x = "Date")
```

```
## Warning: Removed 6 row(s) containing missing values (geom_path).
```

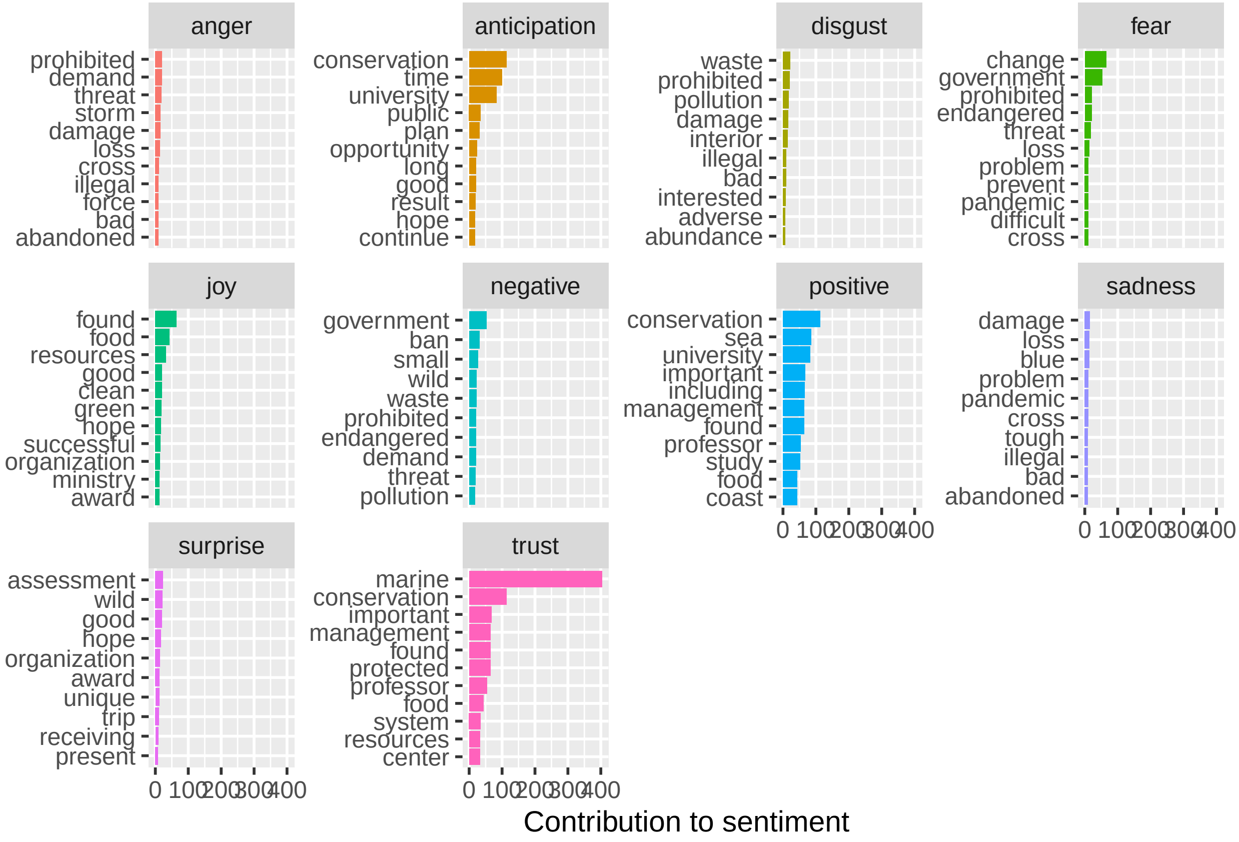


The majority of the words used are words that relate to trust, which, as a marine ecologist myself, I think is really great. There also seems to be a lot of anticipation, that is probably surrounding climate change, rising ocean temperatures, and rising sea levels. I am very interested to see if this changes at all when you increase the time frame that this is over. So I plotted this for a few years of data, and the patterns stays the same over the past 5 years. However I am very interested to know why exactly trust is the most highly ranked word, when there are so many people who now a days don't trust scientists. I wonder if this has to do with where the Nexis library is getting all of these papers from, and anything to do with the nrc groupings. So to look at this I wanted to see what are the main words contributing to the different sentiments. And I found that the main word found in trust was marine, which is probably skewing the data. I am curious why people think that marine is such a trustworthy word. I think it would be interesting to see what this looks like without the word marine.

```
marine_sent_counts <- text_words %>%
        group_by(element_id) %>%
        inner_join(get_sentiments("nrc")) %>%
        group_by(sentiment) %>%
        count(word, sentiment, sort = TRUE) %>%
        ungroup()
```

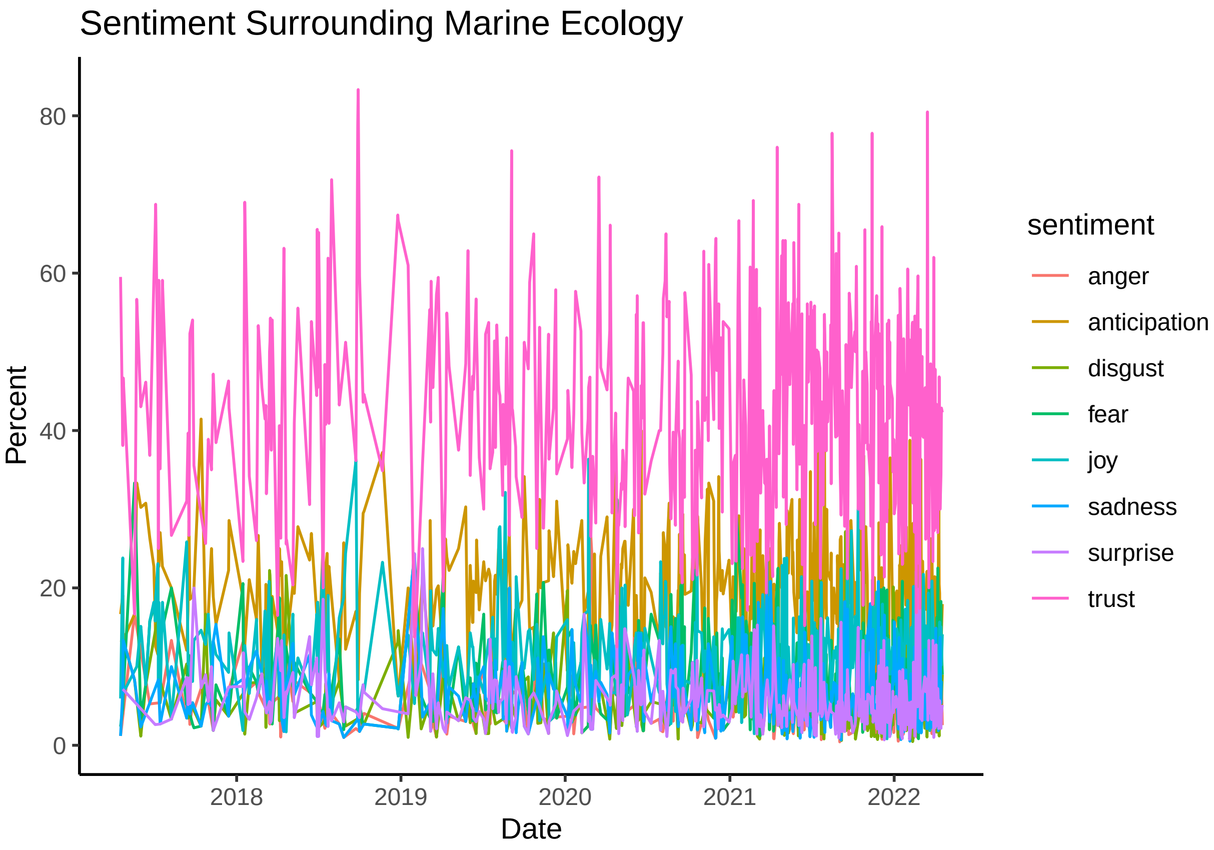
```
## Joining, by = "word"
```

```
marine_sent_counts %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)
```



```
ggplot(data = sent_word_count) +
  geom_line(aes(x = Date, y = percent, fill = sentiment, color = sentiment)) +
  theme_classic() +
  labs(title = "Sentiment Surrounding Marine Ecology",
       y = "Percent",
       x = "Date")
```

```
## Warning: Ignoring unknown aesthetics: fill

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


Sentiment Surrounding Marine Ecology
Percent
80
60
40
20
0
2018
2019
2020
2021
2022
Date
sentiment
anger
anticipation
disgust
fear
joy
sadness
surprise
trust