Week 2 Lab — Exploring word frequencies in New York Times articles related to nuclear fusion

Peter Menzies

4/6/2022

Contents

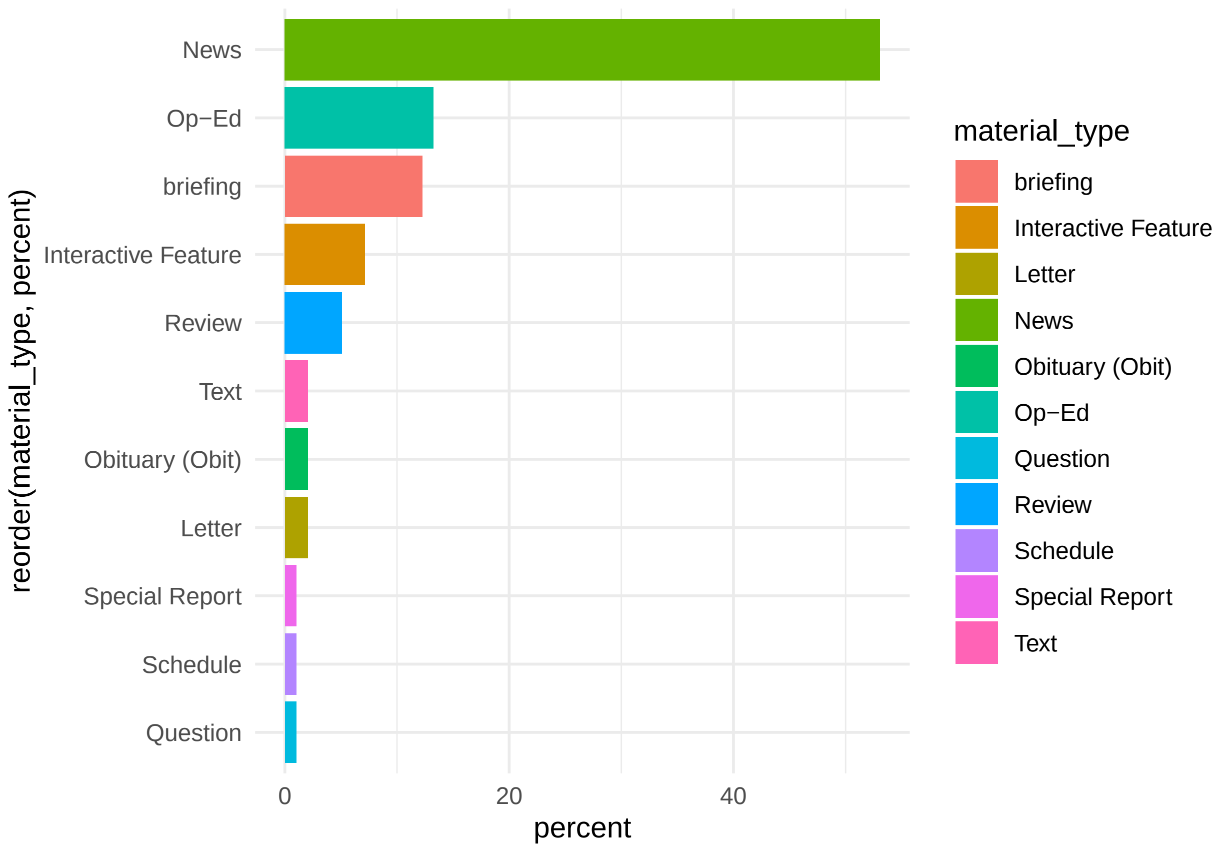
Querying the NYT API

```
if (requery == TRUE)
  {
    term <- "nuclear+fusion" # Need to use + to string together separate words
    begin_date <- "20160101"
    end_date <- "20220411"

    #construct the query url using API operators
    baseurl <- paste0("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=", term,
                      "&begin_date=",begin_date, "&end_date=", end_date,
                      "&facet_filter=true&api-key=", "cqRVqR7OxTbIN7WkN9GJ3DgGsEbMI2cR",
                      sep = "")


    initialQuery <- fromJSON(baseurl)
    maxPages <- round((initialQuery$response$meta$hits[1] / 10) - 1)

    pages <- list()
    for (i in 0:maxPages)
```

```
      {
        nytSearch <- fromJSON(paste0(baseurl, "&page=", i), flatten = TRUE) %>%
          data.frame()
        message("Retrieving page ", i)
        pages[[i+1]] <- nytSearch
        Sys.sleep(6)
    }
    class(nytSearch)

    df <- rbind_pages(pages)

    saveRDS(df, file = "data/nyt_nuclear_fusion.rds")
}

df <- readRDS("data/nyt_nuclear_fusion.rds")
```

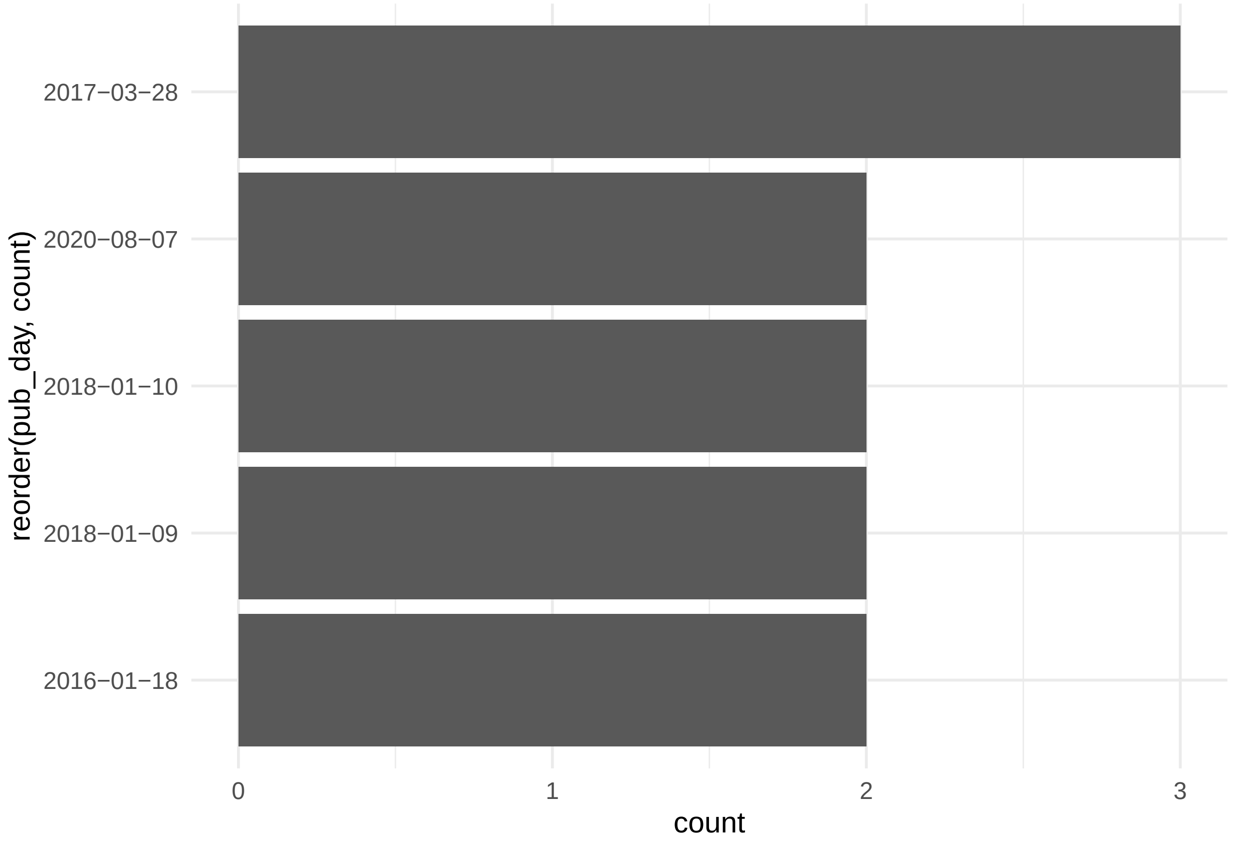
Distribution of article types

```
df %>%
  group_by(response.docs.type_of_material) %>%
  summarize(count = n()) %>%
  mutate(percent = (count / sum(count)) * 100) %>%
  rename(material_type = response.docs.type_of_material) %>%
  ggplot(aes(y = reorder(material_type, percent), x = percent, fill = material_type)) +
  geom_col() +
  theme_minimal()
```



High frequency publication days

```
df %>%
  mutate(pub_day = gsub("T.*", "", response.docs.pub_date)) %>%
  group_by(pub_day) %>%
  summarize(count = n()) %>%
  filter(count >= 2) %>%
  ggplot(aes(y = reorder(pub_day, count), x = count)) +
  geom_col() +
  theme_minimal()
```



Exploring high frequency words

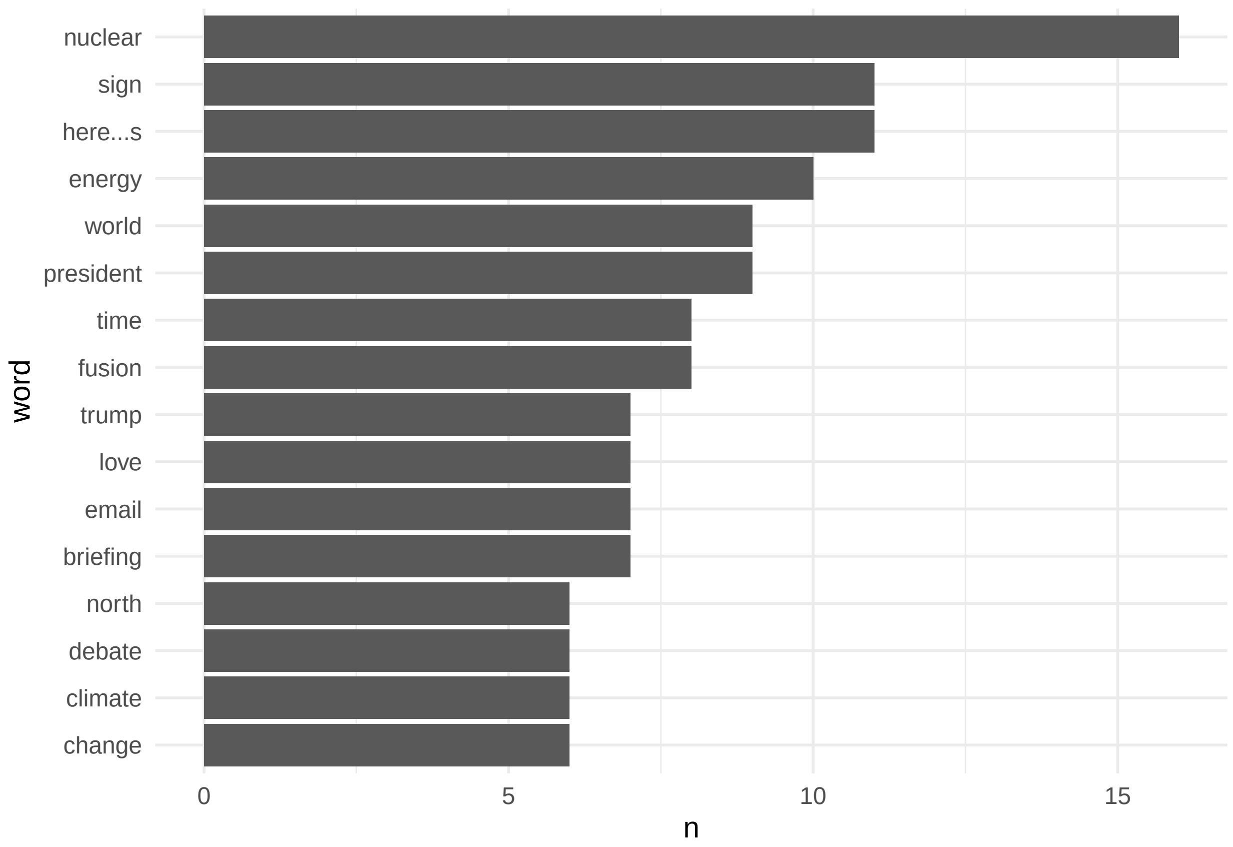
First paragraph

Tokenize and visualize words in first paragraph

```
paragraph <- names(df)[6]
tokenized <- df %>%
  unnest_tokens(word, paragraph)

tokenized <- tokenized %>%
  anti_join(stop_words)
```

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

```
tokenized %>%
  count(word, sort = TRUE) %>%
  filter(n > 5) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = n, y = word)) +
  geom_col() +
  theme_minimal()
```



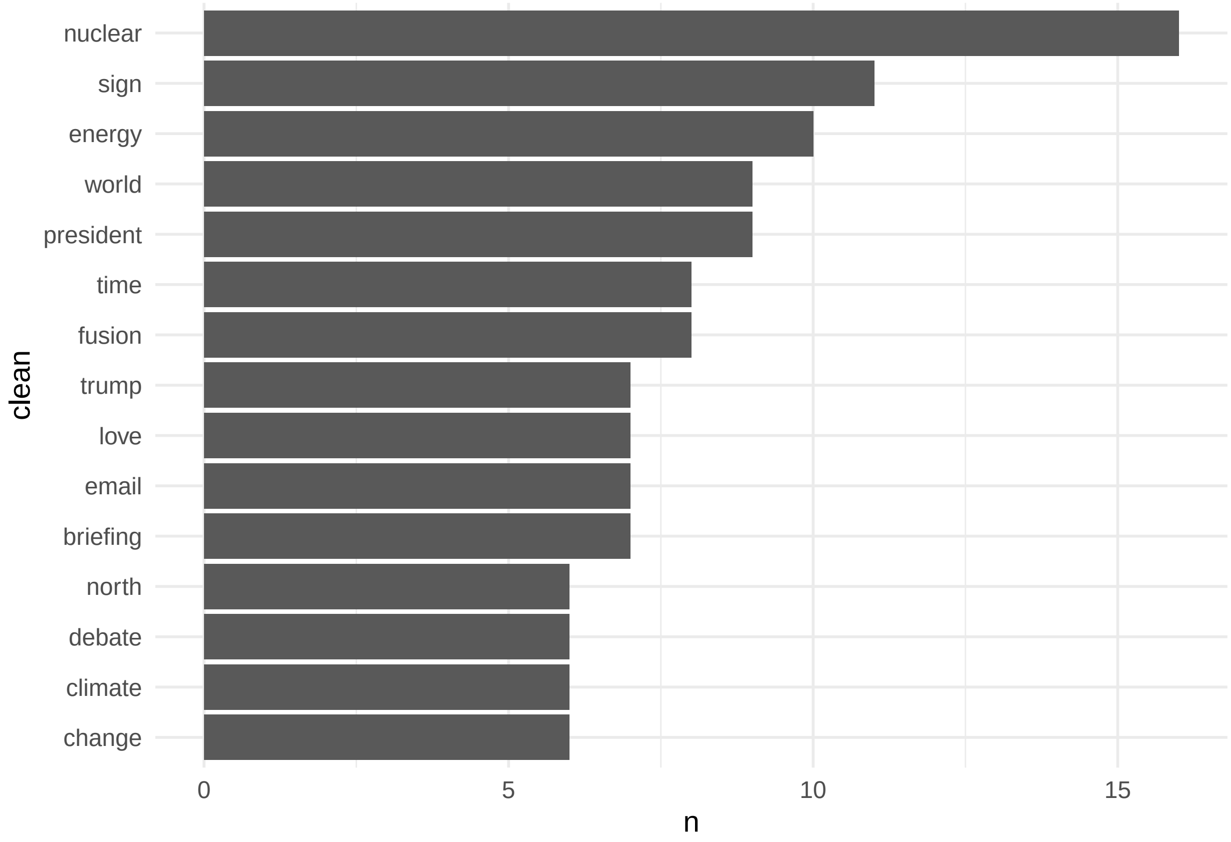
Create and call function to clean up noise and stem potentially important words

```
clean_tokens <- function(df)
  {
    cleaned <- str_remove_all(df$word, "[:digit:]")
    cleaned <- str_replace_all(cleaned, "here.s", "")
    cleaned <- str_replace_all(cleaned, "reactor.", "reactor")
    cleaned <- gsub("'s", "", cleaned)

    return(cleaned)
}
```

```
cleaned_tokens <- clean_tokens(tokenized)

tokenized <- tokenized %>%
  mutate(clean = cleaned_tokens) %>%
  filter(clean != "") %>%
  select(clean)
```

Visualize cleaned up words

```
tokenized %>%
  count(clean, sort = TRUE) %>%
  filter(n > 5) %>%
  mutate(clean = reorder(clean, n)) %>%
  ggplot(aes(x = n, y = clean)) +
  geom_col() +
  theme_minimal()
```

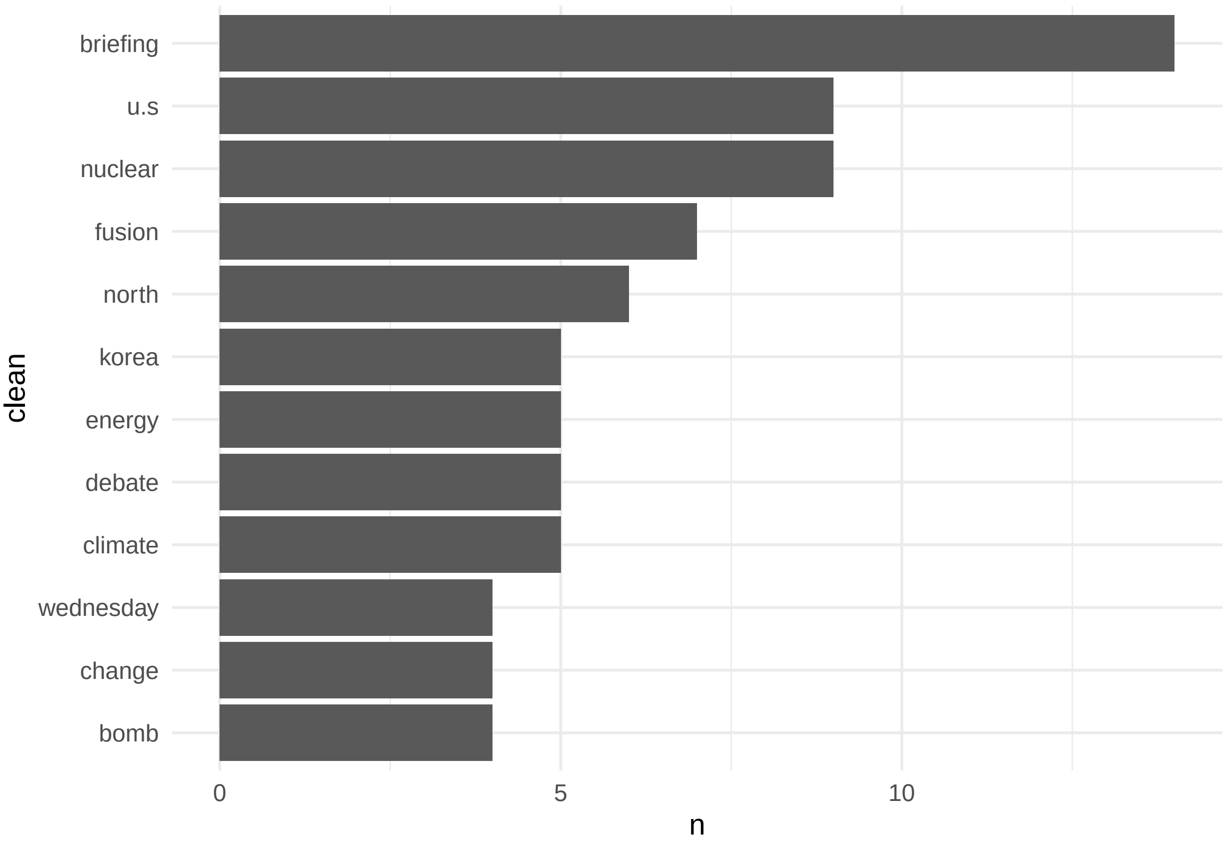


Headlines

```
headlines <- names(df)[21]
tokenized <- df %>%
  unnest_tokens(word, headlines)

tokenized <- tokenized %>%
  anti_join(stop_words)
```

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

```
cleaned_tokens <- clean_tokens(tokenized)

tokenized <- tokenized %>%
  mutate(clean = cleaned_tokens) %>%
  filter(clean != "") %>%
  select(clean)

tokenized %>%
  count(clean, sort = TRUE) %>%
  filter(n > 3) %>%
  mutate(clean = reorder(clean, n)) %>%
  ggplot(aes(x = n, y = clean)) +
  geom_col() +
  theme_minimal()
```



Comparing word frequencies

Generally, high frequency words seem to align for the most part between first paragraphs and headlines. Both distributions illustrate a shortcoming of my search criteria in that many of the articles (unsurprisingly) seem to be related to nuclear weapons and the politics associated with that topic. This is evident from words like “bomb”, “[north] korea”, “trump”—although, “climate” and “energy” show up in both plots, so we’re hitting the mark to some extent. An interesting word that appeared in the paragraph plot and not the headlines plot is “love”—it would be interesting to dig in a little further and investigate why the word “love” is only a bit less frequent than “fusion” which was one of the search terms.