NY Times Text Data in R

Desik Somasundaram

4/6/2022

```
library(jsonlite) #convert results from API queries into R-friendly formats
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks jsonlite::flatten()
## x dplyr::lag()     masks stats::lag()
```

```
library(tidytext) #text data management and analysis
library(ggplot2) #plot word frequencies and publication dates
```

```
#create an object called x with the results of our query ("haaland")
# the from JSON flatten the JSON object, then convert to a data frame
t <- fromJSON("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=haaland&api-key=wtPFWfTfzMOoNH

class(t) #what type of object is x?
```

```
## [1] "list"
```

```
t <- t %>%
  data.frame()


#Inspect our data
class(t) #now what is it?
```

```
## [1] "data.frame"
```

```
dim(t) # how big is it?
```

```
## [1] 10 33
```

```
names(t) # what variables are we working with?
```

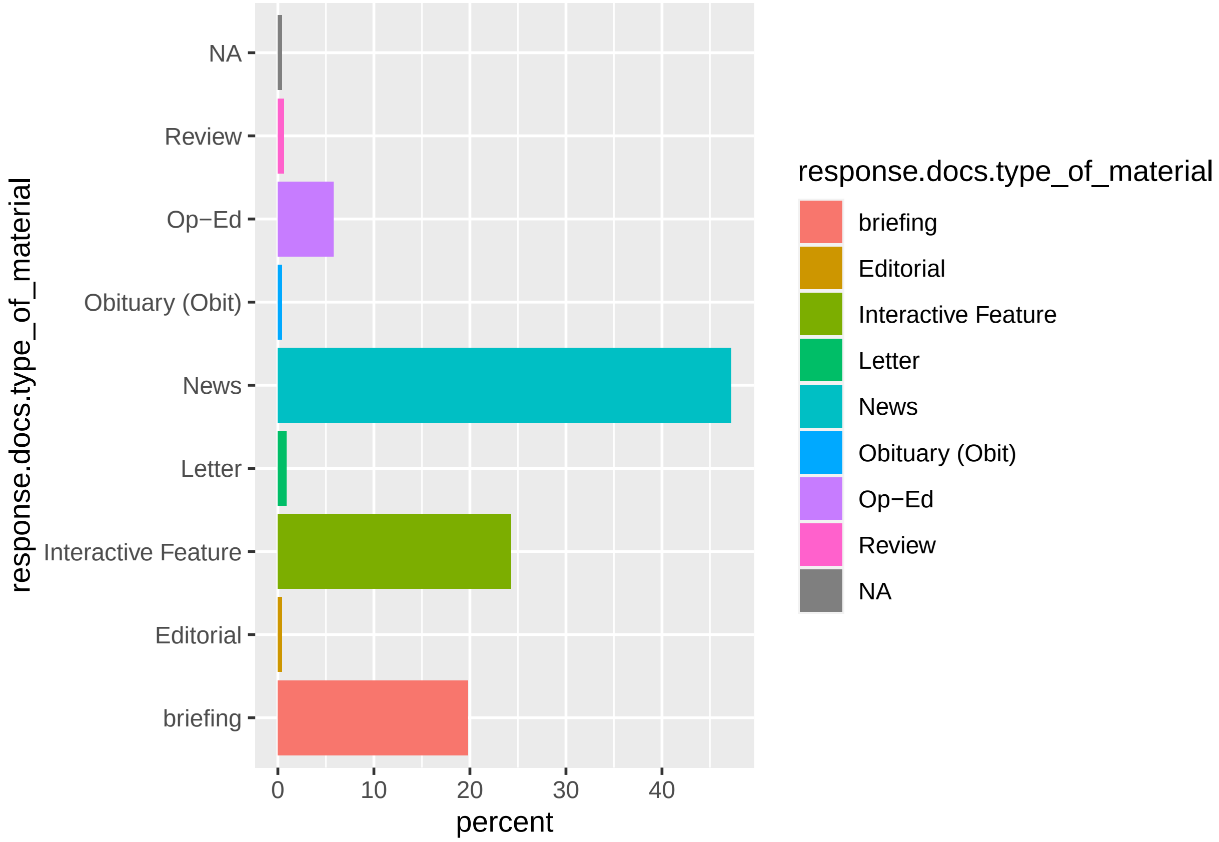
```
##  [1] "status"
##  [2] "copyright"
##  [3] "response.docs.abstract"
##  [4] "response.docs.web_url"
##  [5] "response.docs.snippet"
##  [6] "response.docs.lead_paragraph"
##  [7] "response.docs.source"
##  [8] "response.docs.multimedia"
##  [9] "response.docs.keywords"
## [10] "response.docs.pub_date"
## [11] "response.docs.document_type"
## [12] "response.docs.news_desk"
## [13] "response.docs.section_name"
## [14] "response.docs.type_of_material"
## [15] "response.docs._id"
## [16] "response.docs.word_count"
## [17] "response.docs.uri"
## [18] "response.docs.print_section"
## [19] "response.docs.print_page"
## [20] "response.docs.subsection_name"
## [21] "response.docs.headline.main"
## [22] "response.docs.headline.kicker"
## [23] "response.docs.headline.content_kicker"
## [24] "response.docs.headline.print_headline"
## [25] "response.docs.headline.name"
## [26] "response.docs.headline.seo"
## [27] "response.docs.headline.sub"
## [28] "response.docs.byline.original"
## [29] "response.docs.byline.person"
## [30] "response.docs.byline.organization"
## [31] "response.meta.hits"
## [32] "response.meta.offset"
## [33] "response.meta.time"
```

```
#t <- readRDS("nytDat.rds") #in case of API emergency :)
```

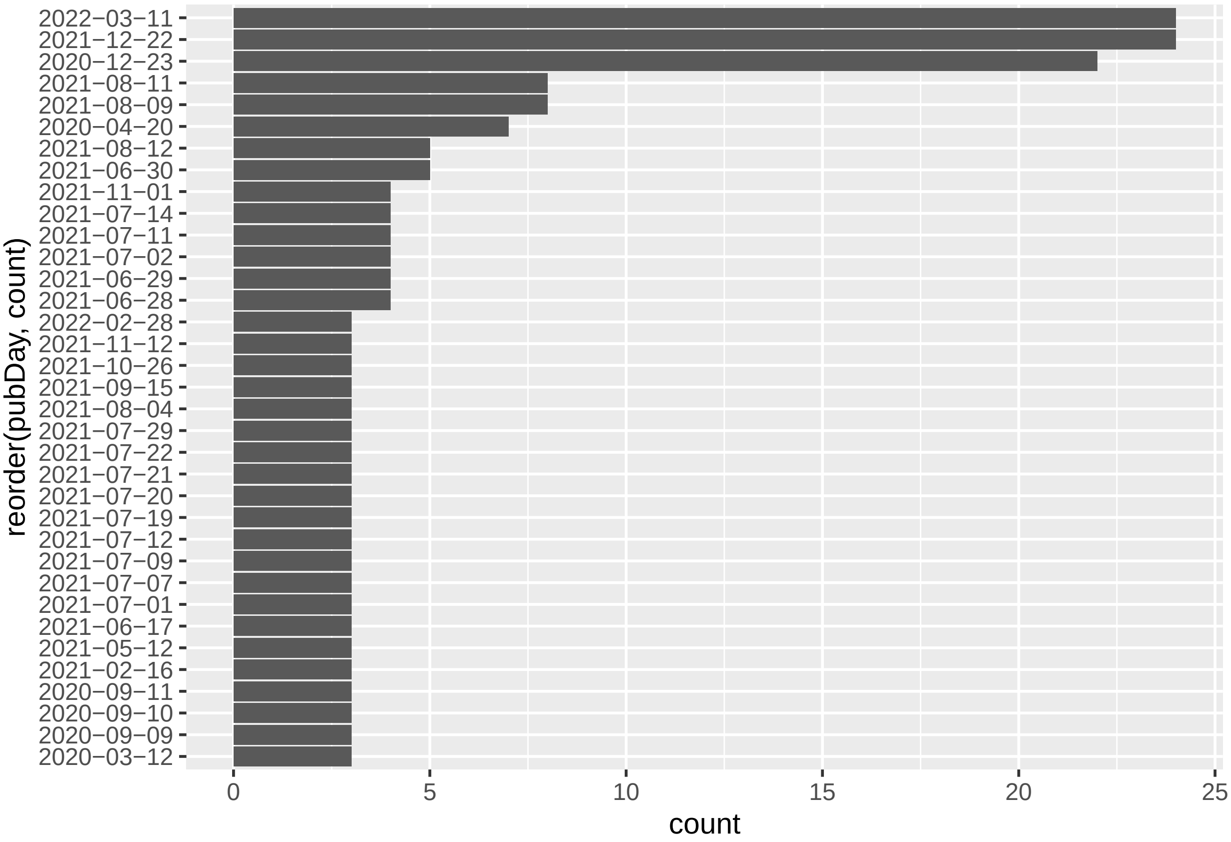
```
term <- "heat+wave+deaths+climate+change" # Need to use + to string together separate words
begin_date <- "20200101"
end_date <- "20220401"

#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=","wtPFWfTfzM0oNHD8wNoIULPLsRhwXNc6", sep="")

#examine our query url
#this code allows for obtaining multiple pages of query results
 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)
#   }
```

```
#need to bind the pages and create a tibble from nytDa
#nytDat <- rbind_pages(pages)
```

```
#nytDat <- nytDat %>%
#  data.frame()

#save(nytDat, file = "nytDat.RData")

load("nytDat.RData")


nytDat %>%
  group_by(response.docs.type_of_material) %>%
  summarize(count=n()) %>%
  mutate(percent = (count / sum(count))*100) %>%
  ggplot() +
  geom_bar(aes(y=percent, x=response.docs.type_of_material, fill=response.docs.type_of_material), stat =
```



```
nytDat %>%
  mutate(pubDay=gsub("T.*","",response.docs.pub_date)) %>%
  group_by(pubDay) %>%
  summarise(count=n()) %>%
  filter(count >= 3) %>%
  ggplot() +
  geom_bar(aes(x=reorder(pubDay, count), y=count), stat="identity") + coord_flip()
```



```
names(nytDat)
```

```
##  [1] "status"
##  [2] "copyright"
##  [3] "response.docs.abstract"
##  [4] "response.docs.web_url"
##  [5] "response.docs.snippet"
##  [6] "response.docs.lead_paragraph"
##  [7] "response.docs.print_section"
##  [8] "response.docs.print_page"
##  [9] "response.docs.source"
## [10] "response.docs.multimedia"
## [11] "response.docs.keywords"
## [12] "response.docs.pub_date"
## [13] "response.docs.document_type"
## [14] "response.docs.news_desk"
## [15] "response.docs.section_name"
## [16] "response.docs.type_of_material"
## [17] "response.docs._id"
## [18] "response.docs.word_count"
## [19] "response.docs.uri"
## [20] "response.docs.subsection_name"
## [21] "response.docs.headline.main"
## [22] "response.docs.headline.kicker"
## [23] "response.docs.headline.content_kicker"
## [24] "response.docs.headline.print_headline"
## [25] "response.docs.headline.name"
## [26] "response.docs.headline.seo"
## [27] "response.docs.headline.sub"
## [28] "response.docs.byline.original"
## [29] "response.docs.byline.person"
## [30] "response.docs.byline.organization"
## [31] "response.meta.hits"
## [32] "response.meta.offset"
## [33] "response.meta.time"
```

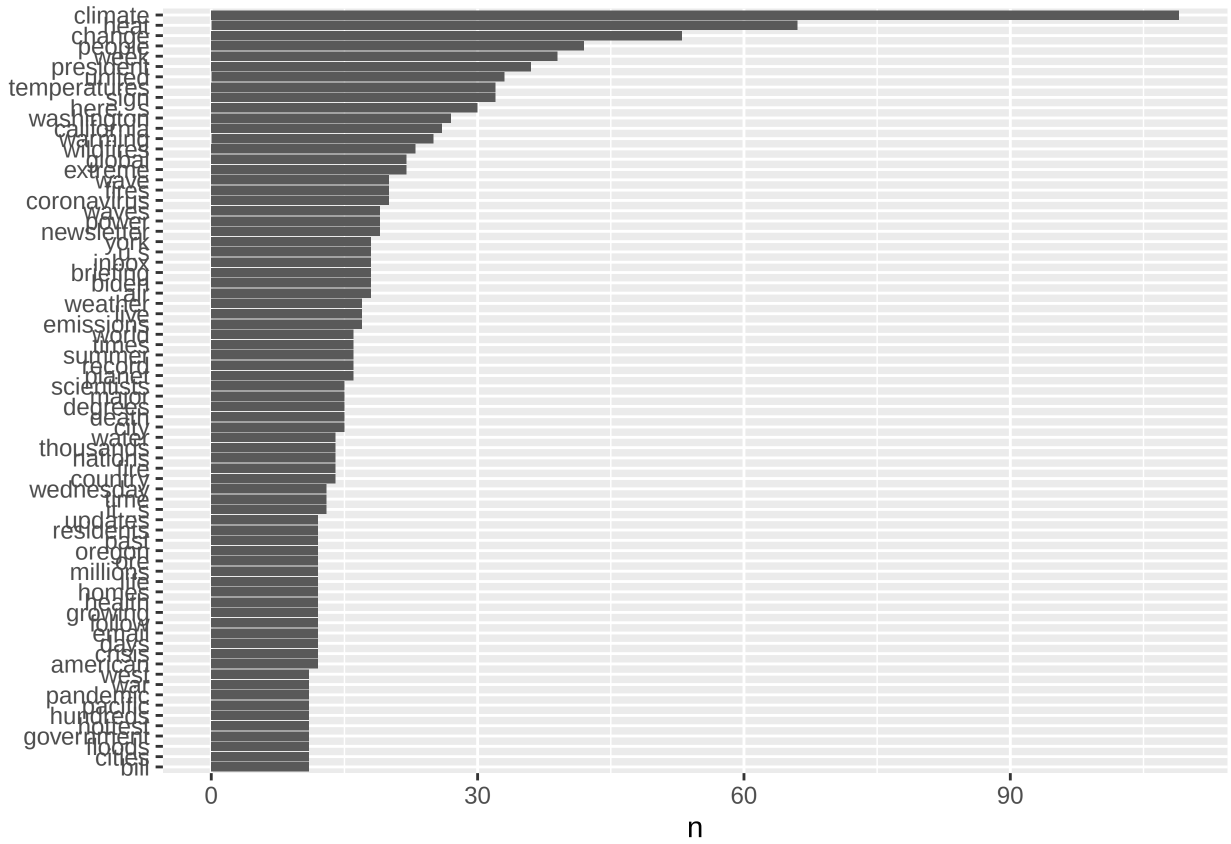
```
paragraph <- names(nytDat)[6] #The 6th column, "response.doc.lead_paragraph", is the one we want here.
tokenized_para <- nytDat %>%
  unnest_tokens(word, paragraph)
```

```
data(stop_words)

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

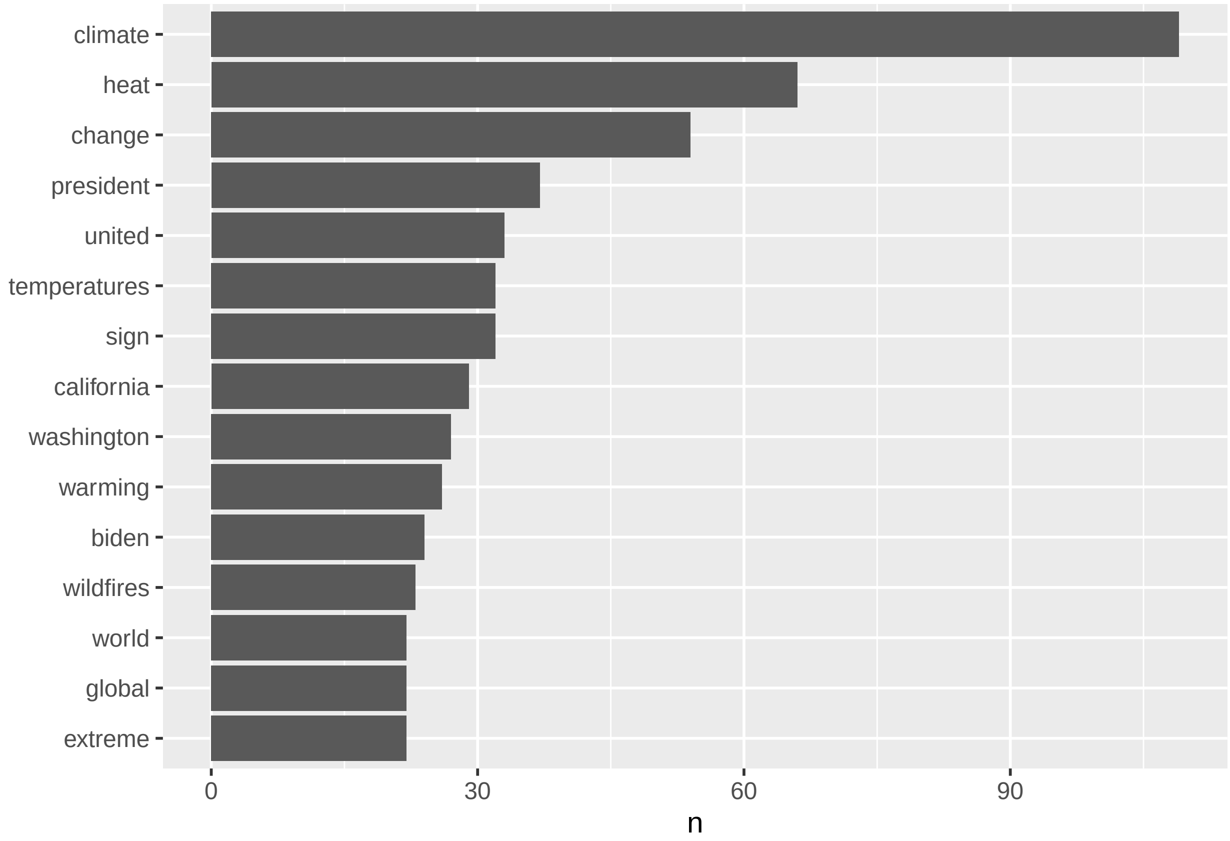
```
## Joining, by = "word"
```

```
tokenized_para %>%
  count(word, sort = TRUE) %>%
  filter(n > 10) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL)
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <80>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’it’s’ in ’mbcsToSbcs’: dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on ’here’s’ in ’mbcsToSbcs’: dot substituted for <99>
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'it's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'here's' in 'mbcsToSbcs': dot substituted for <99>
```



```
#inspect the list of tokens (words)
#tokenized$word

clean_tokens <- str_remove_all(tokenized_para$word, "[:digit:]") #remove all numbers
clean_tokens <- str_remove_all(clean_tokens, "people")
clean_tokens <- str_remove_all(clean_tokens, "here")
clean_tokens <- str_remove_all(clean_tokens, "week")
clean_tokens <- gsub("’s", "", clean_tokens)

tokenized_para$clean <- clean_tokens

#tokenized_para$clean

#remove the empty strings
tib <-subset(tokenized_para, clean!="")

#reassign
tokenized_para <- tib

#try again
tokenized_para %>%
  count(clean, sort = TRUE) %>%
  filter(n > 20) %>%
  mutate(clean = reorder(clean, n)) %>%
  ggplot(aes(n, clean)) +
  geom_col() +
  labs(y = NULL)
```

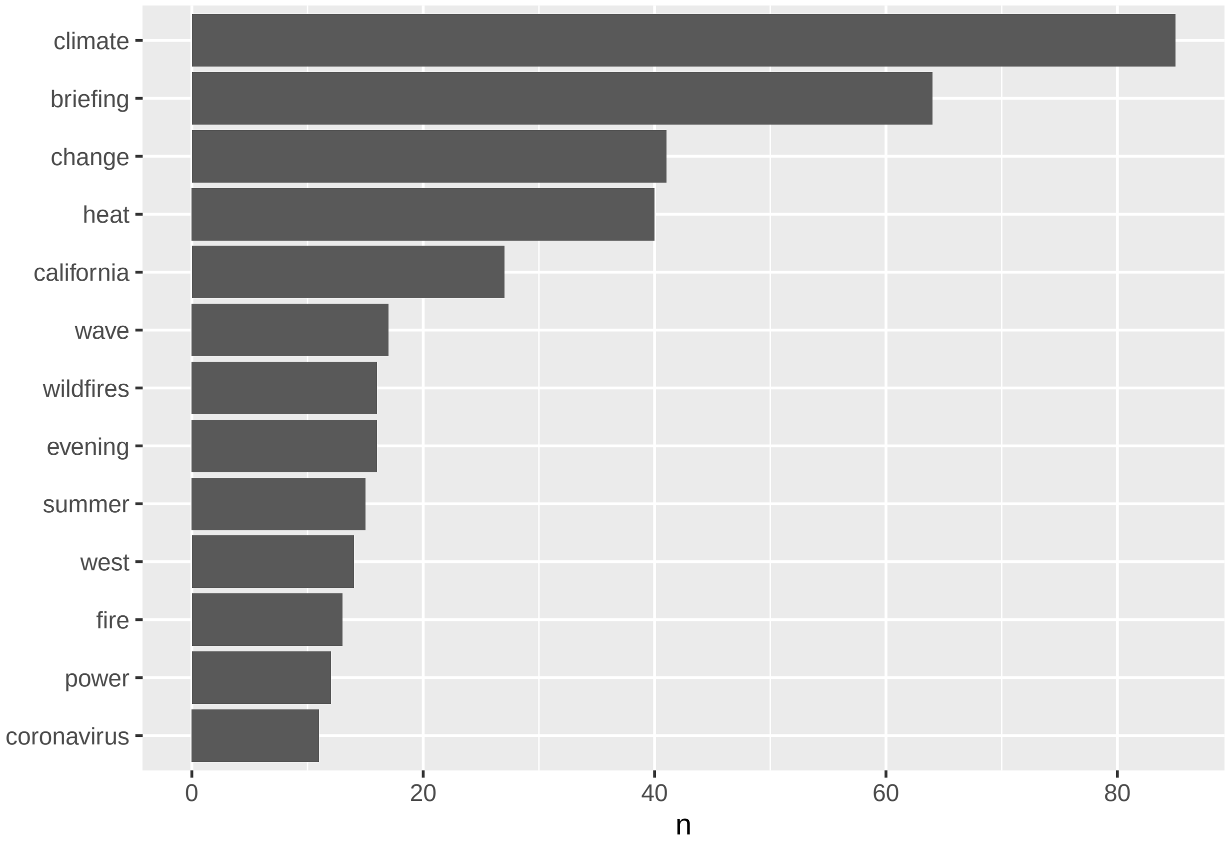


```
headline <- names(nytDat)[21] #The 6th column, "response.doc.headline.main", is the one we want here.
tokenized_hd <- nytDat %>%
  unnest_tokens(word, headline)
```

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

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

```
#inspect the list of tokens (words)
#tokenized$word
days.of.week <- weekdays(Sys.Date()+0:6)

clean_tokens <- str_remove_all(tokenized_hd$word, "[:digit:]") #remove all numbers
clean_tokens <- str_remove_all(clean_tokens, "people")
clean_tokens <- str_remove_all(clean_tokens, "here")
clean_tokens <- str_remove_all(clean_tokens, "week")
for (day in days.of.week){
  clean_tokens <- str_remove_all(clean_tokens, day)
  clean_tokens <- str_remove_all(clean_tokens, tolower(day))
}

clean_tokens <- gsub("’s", "", clean_tokens)

tokenized_hd$clean <- clean_tokens

#tokenized_hd$clean

#remove the empty strings
tib <-subset(tokenized_hd, clean!="")

#reassign
tokenized_hd <- tib

#try again
tokenized_hd %>%
  count(clean, sort = TRUE) %>%
  filter(n > 10) %>%
  mutate(clean = reorder(clean, n)) %>%
  ggplot(aes(n, clean)) +
  geom_col() +
  labs(y = NULL)
```



The comparison of distributions of word frequencies between the first paragraph and headlines had some key similarities but also had noticeable differences. However, it was important and interesting to see that the headlines contain more time-sensitive words such the time of the day, day of the week and season. This required some extra filtering for weekdays to extract a more meaningful word frequency distribution. It was interesting to see that first paragraph distribution contained more larger picture terms and concepts such as president, biden, washington and global.