DATA607_w9 - Web APIs

David Simbandumwe

2021-10-24

Introduction

The New York Times web site provides a rich set of APIs, as described here: https://developer.nytimes.com/apis. For this assignment I selected the top stories api. I will be working with the publication realted data in the science section.

1. created an account
2. configured the authentication
3. Called teh API: https://api.nytimes.com/svc/topstories/v2/science.json
4. Created a data frame: science_df

Import

The first step is calling the API to retrieve a JSON response from the API.

```
# call API
json_url <- "https://api.nytimes.com/svc/topstories/v2/science.json?api-key=ueC6BKg7iaXGRbGD8DALuBFoANFPfalV"
science_df = fromJSON(json_url,
                      simplifyDataFrame = TRUE,
                      flatten = TRUE
            )$results %>%
            as_tibble() %>%
            flatten()
```

Tidy / Transform

The JSON structure is flattened and the article description list and title lists are extracted for further processing.

```
# extract for additional processing
des <- science_df$des_facet %>% unlist()
title <- science_df$title



#clean and tiddy the science_df tibble
science_df <- science_df %>%
  select(section,subsection,title,url,byline, published_date, updated_date) %>%
  mutate (
        published_date = as_date(published_date, tz = NULL, format = NULL),
        updated_date = as_date(updated_date, tz = NULL, format = NULL),
  )
```

Model

The next step is to create a data frame to analyze the description facets and title for each article.

```
# create tibble
des_df <- as.tibble(des)
```

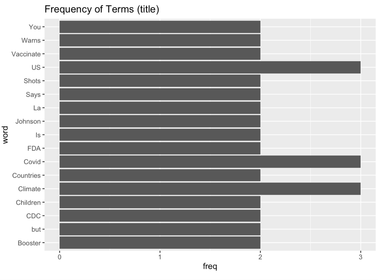
```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
# create dataframe for wordcloud2
des_df <- des_df %>%
  rename(
    word = value
  ) %>%
  group_by(word) %>%
  mutate (
    freq = n()
  ) %>%
  distinct()


title_df <- as.tibble(title)

title_df <- title_df %>%
      separate_rows(
        value,
        convert = TRUE,
        sep = " "
      ) %>%
      rename(
        word = value
      ) %>%
      group_by(word) %>%
      filter(!word %in% c("a","one","to","on","of","in","for","and")) %>%
      mutate (
        word = str_replace_all(word, "[^[:alnum:]]", ""),
        freq = n()
      ) %>%
      distinct()
```

Visualize

Analyse the frequency of key words in the title. Graph the terms by frequency filtering out all the words that only appear once.Create a word cloud that allows us to view the frequency of each word

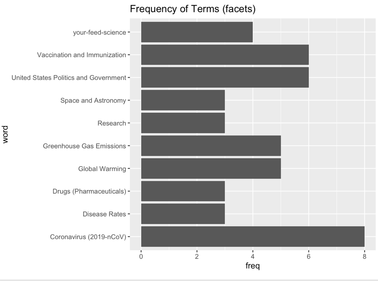
```
title_df %>%
  filter(freq > 1) %>%
  ggplot(aes( y=word, x=freq)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title = "Frequency of Terms (title)" )
```




```
title_df %>%
  wordcloud2(
            color = "random-dark",
            fontFamily = "Helvetica"
           )
```

Analyse the frequency of key words in the facets Graph the terms by frequency filtering out all the words that only appear once.Create a word cloud that allows us to view the frequency of each word

```
des_df %>%
  filter(freq > 2) %>%
  ggplot(aes( y=word, x=freq)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title = "Frequency of Terms (facets)" )
```




```
wordcloud2(des_df,
           color = "random-dark",
           fontFamily = "Helvetica"
          )
```

Conclusions

It is interesting to review the frequency of words in the title vs the facets for each article. As you would expect the facets show less variability than the titles. However it was surprising that the onlye words that appeared mutiltpel times in the titles were US, Covid and Climate. Also it is not surprising that Covid, Global Warming related topics and Politics dominate even the science section. It could be a reflection of how topics like Covid and Global Warming cannot be removed from a political lense.