Data607_HW9

Sean Connin

4/7/21

1. Overview

The NY Times provides an API that enables web-developers and others to access data (1851 to present) associated with the publication. There are currently 10 API's available to the public - for non-commercial use only. They include:

- Archive - past NYT articles for a given month
- Article Search - articles by keyword
- Books - book reviews and The New York Times Best Sellers lists
- Community - comments from registered users on New York Times articles
- Most Popular - comments from registered users on New York Times articles
- Movie Reviews - movie reviews by keyword and opening date
- RSS Feeds - articles ranked on the section fronts
- Semantic - list of people, places, organizations and other locations, entities and descriptors that make up the controlled vocabulary used as metadata by
- Times Newswire - links and metadata for Times' articles
- Top Stories - array of articles currently on the specified sections

In this assignment, I applied the "Article Search" API to access news articles from 2020 highlighting the Adirondack region of New York State. I also took advantage of filtering to limit my search to news articles.

The API returns query results in JSON form. Further steps are required to retrieve and store this data as a dataframe. For example, JSON's data structure (which is hierarchical) must be "flattened" prior to conversion. I used the jasonlite library to accomplish this task. I relied on dplyr to accomplish other aspects of data cleaning and transformation. In this respect, I limited the final data set to entries where Adirondack(s) appeared either in the article headline or lead paragraph.

The following source material was valuable in providing example code and steps to complete this assignment:

- The NY Times Developers Network: https://developer.nytimes.com (https://developer.nytimes.com/)
- Storybench: https://www.storybench.org/working-with-the-new-york-times-api-in-r/ (https://www.storybench.org/working-with-the-new-york-times-api-in-r/)
- Daisung jang: https://daisungjang.com/tutorial/Nytimes_tutorial.html (https://daisungjang.com/tutorial/Nytimes_tutorial.html)

```
#Load libraries

library(tidyverse)
library(jsonlite)
library(magrittr)
library(kableExtra)
library(janitor)
```

2. NY Time API Query

```
#set query parameters to generalize requests

key = "jY4q18VWUjMQPNMBjhrdquQ7AVPxYGZe"

search_terms <- c("Adirondacks")

material<-c("News")

# Query API and identify data volumn (page number)

for(i in 1:length(search_terms)){
    term <- search_terms[i]}

begin_date <- "20000101"
end_date <- "20001231"

url <- paste0("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=",
              term,
              "&fq=type_of_material:",
              material,
              "&begin_date=",
              begin_date,
              "&end_date=",
              end_date,
              "&facet_filter=true&api-key=",
              key, sep="")

#Query hit limited to 10 pages

query <- fromJSON(url)

tot_pages <- round((query$response$meta$hits[1] / 10)-1)

#Create a df for each page and pastes on page number

pages <- list()

for(i in 0:tot_pages){
  nytSearch <- fromJSON(paste0(url, "&page=", i), flatten = TRUE) %>% data.frame()
  message("Retrieving page ", i)
  pages[[i+1]] <- nytSearch
  Sys.sleep(6)
}

#combine dataframes

temp <- rbind_pages(pages)

# Save as CSV after subsetting and renaming columns for clarity

temp%<>%select(!c(status, copyright,response.docs.abstract, response.docs.snippet,response.docs._id,  response.docs.multimed
ia, response.docs.keywords, response.docs.byline.person, response.docs.uri, response.docs.print_section, response.docs.print
_page, response.docs.subsection_name, response.docs.headline.kicker, response.docs.headline.content_kicker, response.docs.he
adline.print_headline, response.docs.headline.name, response.docs.headline.seo, response.docs.headline.sub,response.docs.byl
ine.organization, response.meta.hits, response.meta.offset, response.meta.time))

temp%<>%rename(url=response.docs.web_url, lead = response.docs.lead_paragraph, source =response.docs.source, pub_date = resp
onse.docs.pub_date, type = response.docs.document_type, news_desk = response.docs.news_desk,  section = response.docs.sectio
n_name, material = response.docs.type_of_material, word_count=response.docs.word_count, headline = response.docs.headline.ma
in, author = response.docs.byline.original)

write_csv(temp, "Adks2000.csv")
```

3. Clean and Tidy data

```
#Filter dataset to remove irrelevant articles

Adks<-read_csv("Adks2000.csv")%>%as.data.frame()
View(Adks)

Adks%<>%filter(material != "Paid Death Notice")
Adks%<>%filter(material != "Obituary; Biography")
Adks%<>%filter(news_desk != "classified")

adk1<-Adks%>%filter(str_detect(headline,'Adirondack|Adirondacks'))
adk2<-Adks%>%filter(str_detect(lead,'Adirondack|Adirondacks'))

Adks2000_final<-rbind(adk1, adk2)%>%separate(pub_date, c("date","temp"), sep =" ")%>%select(!temp)%>%clean_names()

#print final df and save to csv

Adks2000_final%>%kbl%>%kable_material(c("striped"))
```

url
https://www.nytimes.com/2000/08/20/us/2000-campaign-president-clinton-maintains-low-profile-his-adirondacks-vacation.html (https://www.nytimes.com/2000/08/20/us/2000-campaign-president-clinton-m
https://www.nytimes.com/2000/03/27/nyregion/acid-rain-law-found-to-fail-in-adirondacks.html (https://www.nytimes.com/2000/03/27/nyregion/acid-rain-law-found-to-fail-in-adirondacks.html)
https://www.nytimes.com/2000/02/21/nyregion/rare-avalanche-kills-one-on-an-adirondack-slope.html (https://www.nytimes.com/2000/02/21/nyregion/rare-avalanche-kills-one-on-an-adirondack-slope.html)
https://www.nytimes.com/2000/11/03/arts/weekend-warrior-alone-in-the-wilds-where-nature-makes-waves.html (https://www.nytimes.com/2000/11/03/arts/weekend-warrior-alone-in-the-wilds-where-nature
https://www.nytimes.com/2000/03/27/nyregion/acid-rain-law-found-to-fail-in-adirondacks.html (https://www.nytimes.com/2000/03/27/nyregion/acid-rain-law-found-to-fail-in-adirondacks.html)
https://www.nytimes.com/2000/02/21/nyregion/rare-avalanche-kills-one-on-an-adirondack-slope.html (https://www.nytimes.com/2000/02/21/nyregion/rare-avalanche-kills-one-on-an-adirondack-slope.html)
https://www.nytimes.com/2000/04/02/weekinreview/march-26-april-1-failure-is-reported-for-clean-air-act.html (https://www.nytimes.com/2000/04/02/weekinreview/march-26-april-1-failure-is-reported-for-cle
https://www.nytimes.com/2000/08/19/nyregion/mrs-clinton-takes-a-busman-s-holiday.html (https://www.nytimes.com/2000/08/19/nyregion/mrs-clinton-takes-a-busman-s-holiday.html)
https://www.nytimes.com/2000/08/20/travel/a-voice-from-the-met-s-first-season.html (https://www.nytimes.com/2000/08/20/travel/a-voice-from-the-met-s-first-season.html)
https://www.nytimes.com/2000/04/10/nyregion/invasive-mussels-turn-up-in-lake-thought-to-be-immune.html (https://www.nytimes.com/2000/04/10/nyregion/invasive-mussels-turn-up-in-lake-thought-to-be-

url
https://www.nytimes.com/2000/06/09/arts/weekend-warrior-stalking-the-bear-bears-optional.html (https://www.nytimes.com/2000/06/09/arts/weekend-warrior-stalking-the-bear-bears-optional.html)
https://www.nytimes.com/2000/07/16/arts/television-radio-saving-the-world-one-sexy-teen-at-a-time.html (https://www.nytimes.com/2000/07/16/arts/television-radio-saving-the-world-one-sexy-teen-at-a-tim

```
write_csv(Adks2000_final, "Adks2000_final")
```

4. Conclusion

The NY Times API provides a reasonably user-friendly interface for accessing publication related data by developers. Users should be aware that query requests will be limited to 10 pages unless query hits are spaced at least 6-seconds apart (per API instructions) and that filter options are limited to those provided by the API.

The query that I constructed produced 12 entries (after data cleaning) for the year 2020. The majority of these entries can be grouped thematically into two categories: 1. environmental; and 2) recreational.