Web APIS- New York Times

Jack Wright

10/23/2020

```
## -- Attaching packages ---------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'jsonlite'

## The following object is masked from 'package:purrr':
##
##     flatten
```

Process

I am using the NY Times API for article searches. I want to search for articles about the New York Jets from the sports desk over the last year. I will be generating the API request programattically.

Creating the request for the GET() function in HTTR:

Load api key

```
location<-"c:\\password-files-for-r\\nytimes_keys.csv"
nytimes_keys<-read.csv(location)
nytimes_keys$api_key
```

```
## [1] "VU21lrYzoKLQAI9tbwOqRQAPdGFmTRpO"
```

```
base_url<-"https://api.nytimes.com/svc/search/v2/articlesearch.json?"
#main query
q<-"q=New+York+Jets"
#&news_desk=sports&begin_date=20200101&end_date=20201023
```

```
#elements of fq
key<-paste0("api-key=",nytimes_keys$api_key)
tag<-paste(q,key,sep="&")

url<-paste0(base_url,tag)
```

Requesting the Data

I will use the GET() function from the httr package and check the status to see if it was successful

```
jets_pull<-GET(url)

http_status(jets_pull)
```

```
## $category
## [1] "Success"
##
## $reason
## [1] "OK"
##
## $message
## [1] "Success: (200) OK"
```

Inspecting the data

The request to the api is structured as a nested named list. I need to find out where the content I am interested in is located.

```
#look at the names
names(jets_pull)
```

```
##  [1] "url"         "status_code" "headers"     "all_headers" "cookies"
##  [6] "content"     "date"        "times"       "request"     "handle"
```

```
#i want the content, but its contents is just raw bytes
glimpse(jets_pull$content)
```

```
##  raw [1:223798] 7b 22 73 74 ...
```

```
#data is in bytes, so convert to text
jets_content<-fromJSON(rawToChar(jets_pull$content))

#after some checks I found where the data I am interested is located
names(jets_content$response$docs)
```

```
##  [1] "abstract"         "web_url"          "snippet"          "lead_paragraph"
##  [5] "print_section"    "print_page"       "source"           "multimedia"
##  [9] "headline"         "keywords"         "pub_date"         "document_type"
## [13] "news_desk"        "section_name"     "subsection_name"  "byline"
## [17] "type_of_material" "_id"              "word_count"       "uri"
```

Convert to Data frame

Since the data is structured as a list, I will convert it to a data frame.

```
df_jets<-data.frame(jets_content$response$docs)
```

#Tidy the Data

The headline column for this data frame is a nested data frame. I will need to unnest it in order to select the main headline.

Then I will create a new dataframe suitable of looking at what my API request returned

```
#unnest headline and put it in its own data frame
df_headline<-unnest(df_jets$headline)
```

```
## Warning: ‘cols‘ is now required when using unnest().
## Please use ‘cols = c()‘
```

```
output<-data.frame("main_headline"=df_headline$main,"abstract"=df_jets$abstract,"date"=df_jets$pub_date)

output%>%
  mutate(ymd=as.Date(date))%>%
  select(-date)
```

```
##                                                                           main_headline
## 1                                   Bill Mathis, a Durable Original Jet, Is Dead at 81
## 2                         After Shutout Loss to Miami, Jets Stand as Only Winless Team
## 3                                  The Jets and Giants Are Both 0-5. Can It Get Worse?
## 4                                                    Jets Cut Ties With Leâ\200\231Veon Bell
## 5                                             Footballâ\200\231s Boo Birds Are All Cooped Up
## 6                                   Jets Start 0-5 for Third Time in Franchise History
## 7  Jets Have a Coronavirus Scare Before a Test Result Turns Out to Be a False Positive
## 8                       Weapons Charge Against Quinnen Williams of the Jets Is Dropped
## 9                                 The Watchable Parts of Thursdayâ\200\231s Broncos-Jets Game
## 10                           One Depleted Team Played Well Sunday (Hint: Not the Jets)
##
## 1                                           A versatile running back, he spent his entire 10-year care
## 2                                             Ryan Fitzpatrick threw three touchdowns for the Dolphins
## 3                                                        This may not be the worst year in New York fo
## 4                                                          The Jets released the running back, their r
## 5  N.F.L. fans in the Northeast, lusty booers in normal times, have had to watch their teamsâ\200\23
## 6                                               Arizona quarterback Kyler Murray scored two touchdowns
## 7                                   On Friday evening, the team reported that the whole squad had ulti
## 8                                            The charge against Williams, a defensive lineman for the
## 9                                                                                A sloppy contest betw
## 10                                      The San Francisco 49ers lost two stalwarts from their defens
##           ymd
## 1  2020-10-22
## 2  2020-10-18
## 3  2020-10-14
## 4  2020-10-14
## 5  2020-10-18
## 6  2020-10-11
## 7  2020-10-09
## 8  2020-10-05
## 9  2020-10-02
## 10 2020-09-20
```

#Conclusions

My pull was only for the first page of results of jets articles (the most recent). I could create a function that allows me to add a pagination facet, allowing me to cycle through the results pages and pull more data. I could have also added more facets to my data frame, like only pulling from the sports desk.