Data 607 - Week 9 Assignment

Adam Douglas

10/24/2018

Assignment

Our assignment is to choose one of the New York Times APIs, construct an interface in R to read in the JSON data, and transform it to an R dataframe.

For this exercise, I chose the NY Times Bestseller Lists via the books API[1].

API

To get data from the API we construct a URL with the parameters we need to pass:

```
# The NY Times best-seller list API URL
baseURL <- "https://api.nytimes.com/svc/books/v3/lists.json"

# The specific list type. We're going to look at hard-cover fiction
listing <- "hardcover-fiction"

# The date for the very first listing (per the API)
asOf <- "2008-06-08"

# Now we build the final URL. Our API key is stored as an option
url <- paste(baseURL, "?", "&api-key=", getOption("NYTimesAPIKey"),
             "&list=", listing, "&date=", asOf, sep = "")
```

Now that we have contructed the proper URL, we can get the JSON reply from the API.

```
# Get the JSON from the URL
res <- fromJSON(url)

# Get the results in a data frame
books <- res$results
```

The results

Looking at the data frame that the `fromJSON` function gave us, we see that in some cases the variables we were given were actually lists.

```
# Check the structure
class(books$book_details)
```

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

The lists actually seem to contain data frames with additional details about the book.

```
class(books$book_details[[1]])
```

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

[1] http://developer.nytimes.com/books_api.json

```
books$book_details[[1]]
```

```
##       title
## 1 ODD HOURS
##                                                                                description
## 1 Odd Thomas, who can communicate with the dead, confronts evil forces in a California coastal town.
##         contributor        author contributor_note price age_group
## 1 by Dean R. Koontz Dean R Koontz                      27
##   publisher primary_isbn13 primary_isbn10
## 1    Bantam  9780553807059     0553807056
```

The same sort of thing applies to the *reviews* column:

```
class(books$reviews[[1]])
```

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

```
books$reviews[[1]]
```

```
##   book_review_link first_chapter_link sunday_review_link
## 1
##   article_chapter_link
## 1
```

Looking at the *isbns* column, however, there appears to be several observations in each data frame (and some with none), meaning that one book can have several ISBNs. This makes sense because books may have different covers, have international editions, large-print, etc. So, we will leave these nested.

```
class(books$isbns[[1]])
```

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

```
books$isbns[[1]]
```

```
## data frame with 0 columns and 0 rows
```

```
books$isbns[[2]]
```

```
##       isbn10        isbn13
## 1 0316068055 9780316068055
## 2 0316068047 9780316068048
## 3 0316043044 9780316043045
## 4 0316128651 9780316128650
## 5 0316218502 9780316218504
## 6 0316218510 9780316218511
```

For the first two lists, we can extract the data frames into one larger data frame (since there is only one observation in each)

```
# First the details
details <- bind_rows(books$book_details)

# Then the reviews
reviews <- bind_rows(books$reviews)
```

As expected, these have the same number of records as our main data frame

```
dim(details)
```

```
## [1] 20 10
```

```
dim(reviews)
```

```
## [1] 20  4
```

Now we can combine those data frames with the main frame, combining the data on each book into a more convenient format.

```
books <- bind_cols(books, details, reviews)
```

Now, with the data in a convenient format, we can do whatever analysis we wish or, if required, return to the API and retreive more data:

```
################################################
# Get all 52 weekly best seller lists into one  #
# single data frame for analysis                #
################################################

# Get a character vector of 52 weeks, as the list is pulished every week
dates <- rep(ymd("2008-06-08"),52)

for (i in 2:52){
    dates[i] <- dates[i-1] + weeks(1)
}

# Create a list for our result data frames to be stored in
staging <- list(data.frame(c(1,1)))

# Get the API data for each week
for (x in 1:52){
  # Get the week
  asOf <- as.character(dates[x])

  # Build the URL for the week
  url <- paste(baseURL, "?", "&api-key=", getOption("NYTimesAPIKey"),
               "&list=", listing, "&date=", asOf, sep = "")

  # Get the data and parse it
  res <- fromJSON(url)
  staging[[x]] <- res$results

  # A small pause
  Sys.sleep(1)

}

# Combine the data frames
allBooks <- bind_rows(staging)

# Now fix the data as we did above by pulling out the sublists
details <- bind_rows(allBooks$book_details)
reviews <- bind_rows(allBooks$reviews)
allBooks <- bind_cols(allBooks, details, reviews)
```

Now with a year's worth of best sellers, we can do whatever analysis we want. For example, looking at which publisher had the most best-sellers each week:

```r
knitr::kable(allBooks %>% group_by(bestsellers_date, publisher) %>% tally() %>% top_n(1,n), col.names =
          caption="NY Times Bestseller List (Hardcover Fiction) Top Publishers") %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE, position = "left")
```

Table 1: NY Times Bestseller List (Hardcover Fiction) Top Publishers

Week	Publisher	# Books on Bestseller List
2008-05-24	Harper	3
2008-05-24	Little, Brown	3
2008-05-31	Harper	3
2008-05-31	Little, Brown	3
2008-06-07	Putnam	4
2008-06-14	Little, Brown	3
2008-06-21	Little, Brown	3
2008-06-21	Putnam	3
2008-06-21	St. Martin's	3
2008-06-28	Little, Brown	3
2008-06-28	St. Martin's	3
2008-07-05	Simon & Schuster	3
2008-07-12	Simon & Schuster	4
2008-07-19	Putnam	3
2008-07-19	Simon & Schuster	3
2008-07-26	Putnam	3
2008-07-26	Simon & Schuster	3
2008-08-02	Little, Brown	2
2008-08-02	Morrow	2
2008-08-02	Putnam	2
2008-08-02	Simon & Schuster	2
2008-08-02	St. Martin's	2
2008-08-09	St. Martin's	4
2008-08-16	Putnam	3
2008-08-16	St. Martin's	3
2008-08-23	Putnam	4
2008-08-23	St. Martin's	4
2008-08-30	Putnam	4
2008-09-06	Grand Central	3
2008-09-13	Grand Central	2
2008-09-13	Simon & Schuster	2
2008-09-20	Del Rey	2
2008-09-27	Knopf	2
2008-09-27	Putnam	2
2008-09-27	Simon & Schuster	2
2008-09-27	William Morrow	2
2008-10-04	Grand Central	2
2008-10-04	Knopf	2
2008-10-04	Morrow	2
2008-10-04	Putnam	2
2008-10-11	Putnam	3
2008-10-18	Little, Brown	2
2008-10-18	Morrow	2
2008-10-18	Putnam	2
2008-10-25	Little, Brown	3
2008-11-01	Little, Brown	3
2008-11-08	Grand Central	3
2008-11-08	Little, Brown	3
2008-11-15	Grand Central	3
2008-11-15	Little, Brown	3
2008-11-22	Grand Central	3
2008-11-22	Knopf	3
2008-11-22	Little, Brown	3
2008-11-29	Grand Central	3
2008-11-29	Little, Brown	3