

Best Seller Books

19th of June 2020

Data Thieves Project by
Pancakes:

- Milena
- Nektarios
- Sabrina

OUR WORKFLOW

01

Our topic

03

Web scraping

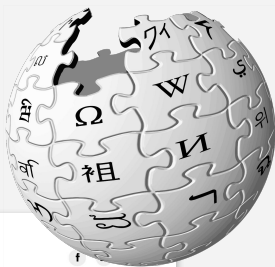
02

New York Times API

04

Data wrangling

The New York Times



BOOKS


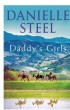

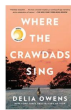

The New York Times Best Sellers

Authoritatively ranked lists of books sold in the United States, sorted by format and genre.

FICTION ▾ | NONFICTION ▾ | CHILDREN'S ▾ | MONTHLY LISTS ▾

< June 28, 2020 >











Combined Print & E-Book Fiction >

1		2		3		4		5	
<p>NEW THIS WEEK THE SUMMER HOUSE by James Patterson and Brendan DuBois</p> <p>Jeremiah Cook, a veteran and former N.Y.P.D. cop, investigates a mass murder near a lake in Georgia.</p> <p>BUY -</p>		<p>NEW THIS WEEK DADDY'S GIRLS by Danielle Steel</p> <p>After a California rancher's sudden death, his three daughters discover things they did not know about their father.</p> <p>BUY -</p>		<p>NEW THIS WEEK TOM CLANCY, FIRING POINT by Mike Maden</p> <p>When an old friend is killed during the bombing of a Barcelona cafe, Jack Ryan Jr. searches for those responsible.</p> <p>BUY -</p>		<p>80 WEEKS ON THE LIST WHERE THE CRAWDADS SING by Delia Owens</p> <p>In a quiet town on the North Carolina coast in 1969, a young woman who survived alone in the marsh becomes a murder suspect.</p> <p>BUY -</p>		<p>2 WEEKS ON THE LIST THE VANISHING HALF by Brit Bennett</p> <p>The lives of twin sisters who run away from a Southern black community at age 16 diverge as one returns and the other takes on a different racial identity but their fates intertwine.</p> <p>BUY -</p>	

01

Our topic

- Choosing the topic
- At a certain point just checked what kind of APIs are available
- Then each of us tried to get the data via the NYT API

 <p>Archive API</p> <p>Get all NYT article metadata for a given month.</p>	 <p>Article Search API</p> <p>Search for New York Times articles.</p>	 <p>Books API</p> <p>Get NYT Best Sellers Lists and lookup book reviews.</p>	 <p>Community API</p> <p>Get user comments. (BETA)</p>	 <p>Geo API</p> <p>Geographic linked data. (DEPRECATED)</p>
 <p>Most Popular API</p> <p>Popular articles on NYTimes.com.</p>	 <p>Movie Reviews API</p> <p>Search for movie reviews.</p>	 <p>RSS Feeds</p> <p>NYT RSS section feeds.</p>	 <p>Semantic API</p> <p>Get semantic terms (people, places, organizations, and locations).</p>	 <p>Times Tapes API</p> <p>NYT controlled vocabulary.</p>

02

New York Times API

- The API giving the data about the best seller books comes with a lot of different categories
- Important to think about the limited requests (time.sleep)

Developers

HomeAPIsCovid-19 Data

Books API

BOOKS API

Overview

PATHS

/lists.json GET

> /lists/...

/reviews.json GET

GET /lists.json

Get Best Sellers list. If no date is provided returns the latest list.

HTTP request

`https://api.nytimes.com/svc/books/v3/lists.json`

Query Parameters

list (required)	string Default value <code>"hardcover-fiction"</code> Default value <code>"hardcover-fiction"</code>
bestsellers-date	string <code>matches ^\d{4}-\d{2}-\d{2}\$</code> <code>matches ^\d{4}-\d{2}-\d{2}\$</code> YYYY-MM-DD The week-ending date for the sales reflected on list-name. Times best sellers lists are compiled using available book sale data. The bestsellers-date may be significantly earlier than published-date. For additional information, see the explanation at the bottom of any best-seller list page on NYTimes.com (example: Hardcover Fiction, published Dec. 5 but reflecting sales to Nov. 29).

```
# set timer to avoid rate limit  
time.sleep(6)
```

Books and book series

(December 1, 2007 – January 1, 2020).

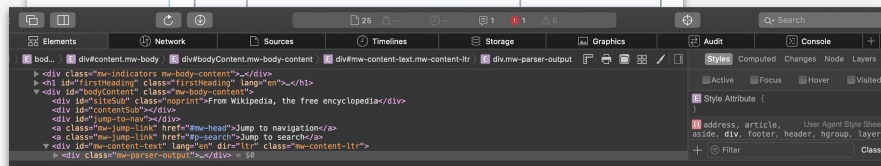
Naruto is classified as a TV series (above) as it is more popular as an anime than by that of its TV adaptation *Game of Thrones*.

Rank	Page	Views in millions
1	Harry Potter	77
2	Fifty Shades of Grey	49
3	A Song of Ice and Fire	48
4	The Hunger Games	46
5	Bible	34
6	The Great Gatsby	32
7	The Lord of the Rings	30
8	Nineteen Eighty-Four	28
9	A Game of Thrones	27
10	Kama Sutra	25
11	To Kill a Mockingbird	24
12	Alice's Adventures in Wonderland	23
12	Les Misérables	23
14	The Handmaid's Tale	22
14	Watchmen	22
16	The Winds of Winter	21
17	Pride and Prejudice	20
17	Harry Potter and the Deathly Hallows	20

03

Web scraping

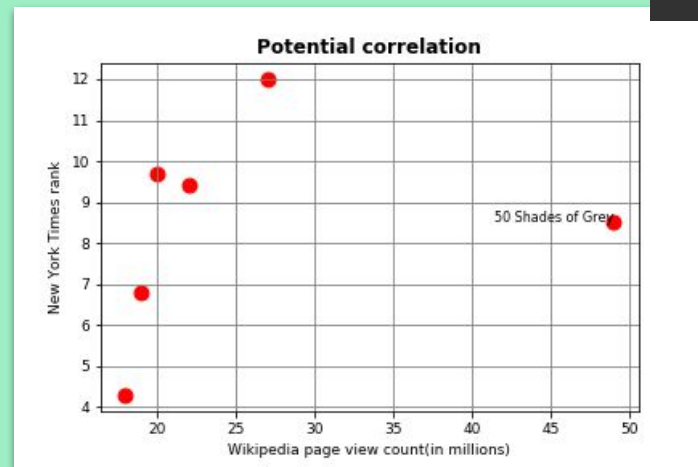
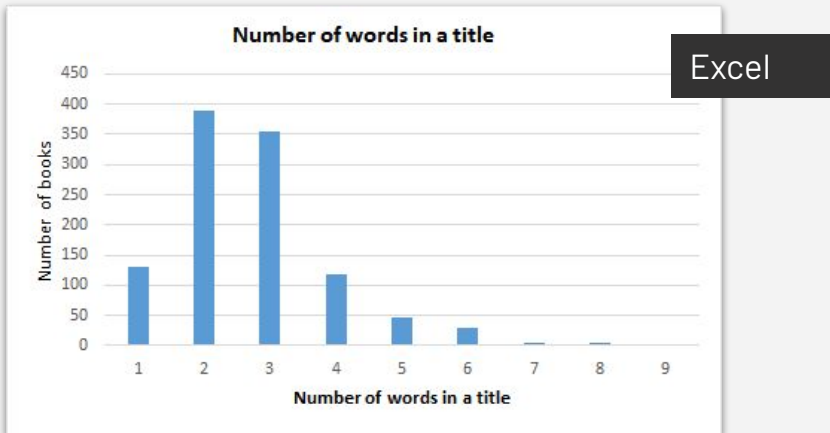
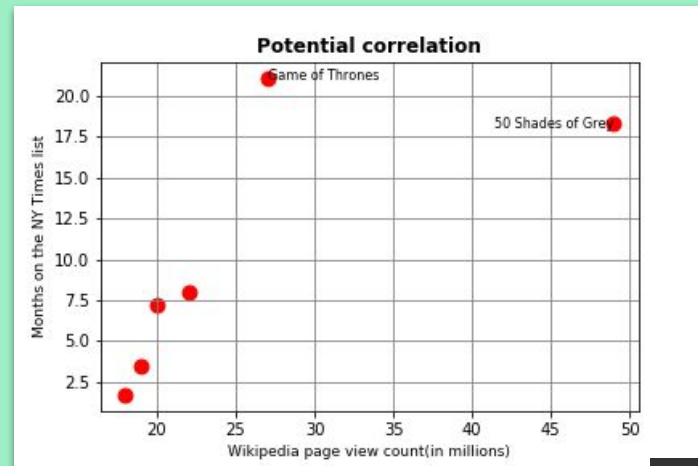
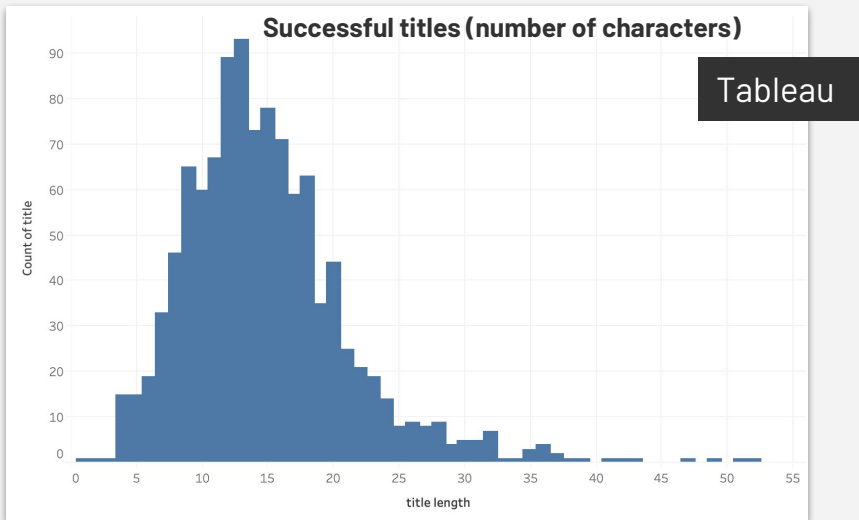
- We were scared of web scraping and wanted to avoid it
- Then decided to try it, as it was only one table (“that can’t be that hard”)



04

Data cleaning

	Book	Author	Wiki_Rank	NYT_Rank(avg)	Months_on_NYT_list(avg)	From	Wiki_Page_Visits
0	FIFTY SHADES OF GREY	E L James	2	8.5	18.3	2012-04-01	49
1	A GAME OF THRONES	George RR Martin	9	12.0	21.1	2011-05-01	27
2	THE HANDMAID'S TALE	Margaret Atwood	14	9.4	8.0	2017-03-01	22
3	THE GIRL WITH THE DRAGON TATTOO	Stieg Larsson	17	9.7	7.2	2011-03-01	20
4	INFERNO	Dan Brown	21	6.8	3.5	2013-06-01	19
5	IT	Stephen King	22	4.3	1.7	2017-10-01	18



Matplotlib

**Thank you
very much!**