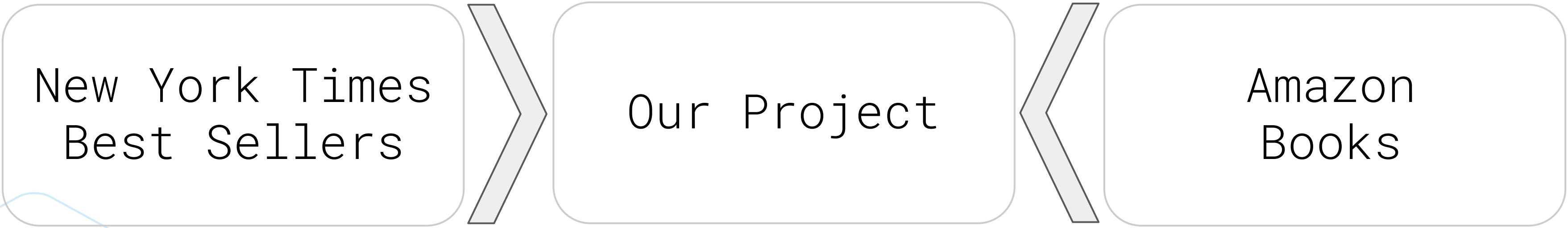
DATA THIEVES PROJECT

Pedro Afonso, Rennê Cirqueira, Thais Ternus

Data Analytics
Part-time Course

GOAL

HELP A PERSON CHOOSING A BOOK FROM
“THE BEST OF THE BEST”



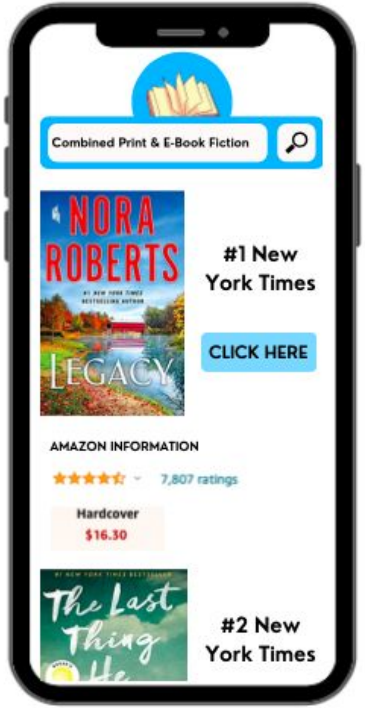
New York Times Best Sellers

Our Project

Amazon Books

RIGHT BOOK

PRODUCT



Combined Print & E-Book Fiction

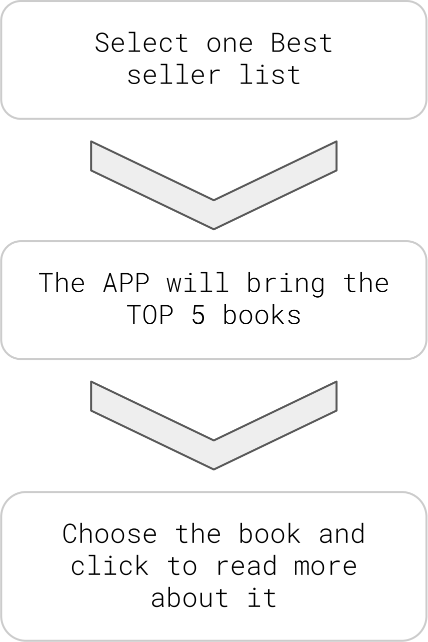
#1 New York Times

AMAZON INFORMATION

7,807 ratings

Hardcover
$16.30

#2 New York Times



Select one Best seller list

The APP will bring the TOP 5 books

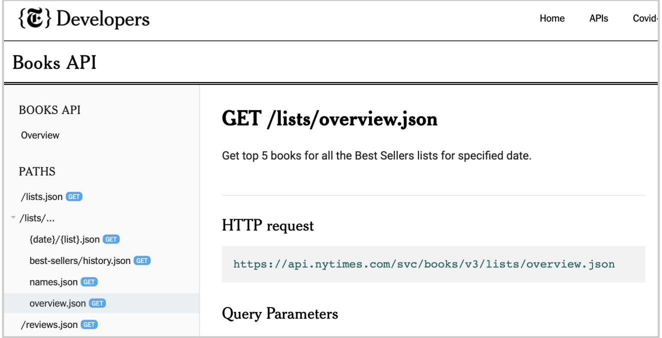
Choose the book and click to read more about it

IRON HACK

API CONNECTION

METHODOLOGY

- ➔ Best Sellers API comes with several functions
- ➔ “Overview” retrieves the top 5 for all categories
- ➔ Right now, the API returns more than 50 books in at least 10 categories
- ➔ All the categories were analysed to be included on the project



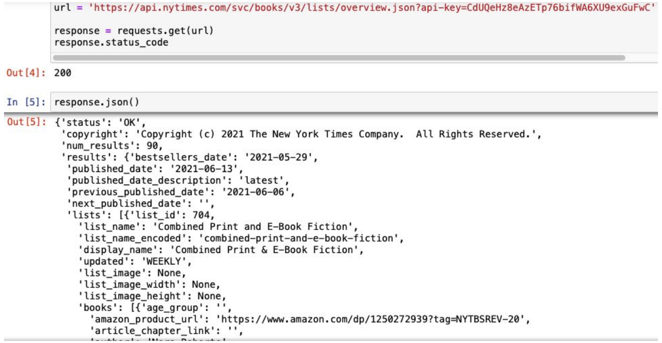
GET /lists/overview.json

Get top 5 books for all the Best Sellers lists for specified date.

HTTP request

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

Query Parameters



```
url = 'https://api.nytimes.com/svc/books/v3/lists/overview.json?api-key=CdUQeHz8eAzETp76bifWA6XU9exGuFwC'

response = requests.get(url)
response.status_code
```

```
Out[4]: 200
```

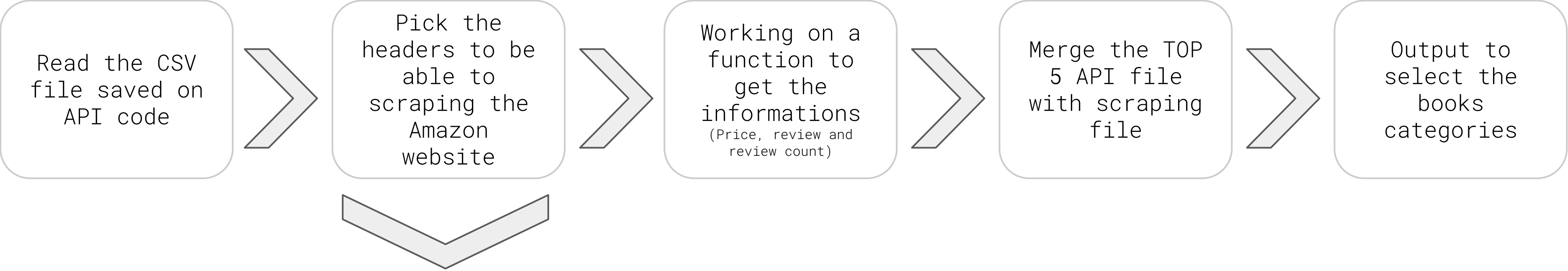
```
In [5]: response.json()
```

```
Out[5]: {'status': 'OK',
         'copyright': 'Copyright (c) 2021 The New York Times Company.  All Rights Reserved.',
         'num_results': 90,
         'results': {'bestsellers_date': '2021-05-29',
          'published_date': '2021-06-13',
          'published_date_description': 'latest',
          'previous_published_date': '2021-06-06',
          'next_published_date': '',
          'lists': [{'list_id': 704,
            'list_name': 'Combined Print and E-Book Fiction',
            'list_name_encoded': 'combined-print-and-e-book-fiction',
            'display_name': 'Combined Print & E-Book Fiction',
            'updated': 'WEEKLY',
            'list_image': None,
            'list_image_width': None,
            'list_image_height': None,
            'books': [{'age_group': '',
              'amazon_product_url': 'https://www.amazon.com/dp/1250272939?tag=NYTBSREV-20',
              'article_chapter_link': '',
```

WEB SCRAPING

amazon

METHODOLOGY



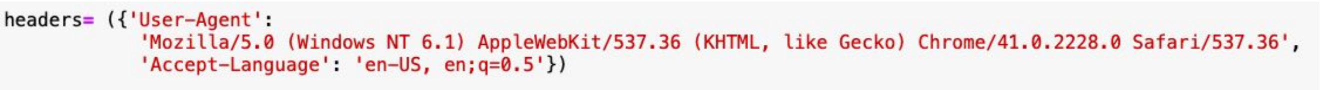
Read the CSV file saved on API code

Pick the headers to be able to scraping the Amazon website

Working on a function to get the informations (Price, review and review count)

Merge the TOP 5 API file with scraping file

Output to select the books categories

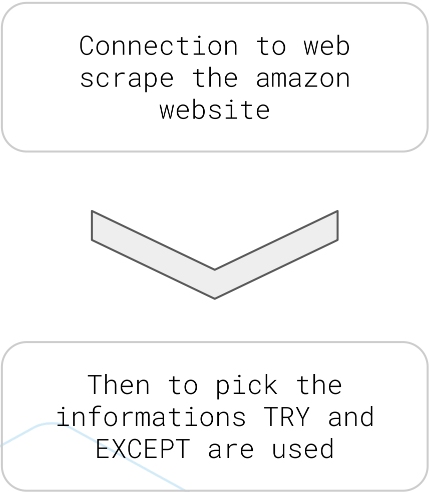


```
headers= ({'User-Agent':
            'Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2228.0 Safari/537.36',
            'Accept-Language': 'en-US, en;q=0.5'})
```

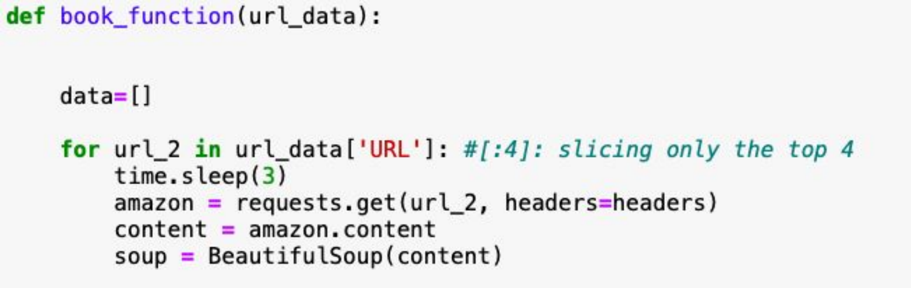
IRON HACK

WEB SCRAPING

FUNCTION METHODOLOGY

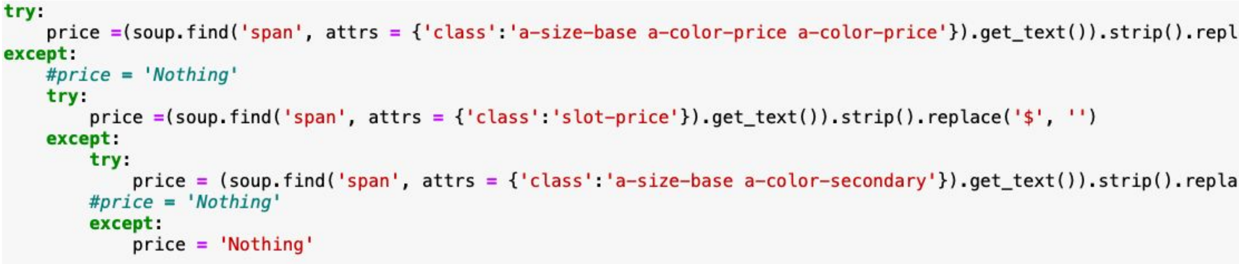

Connection to web scrape the amazon website

Then to pick the informations TRY and EXCEPT are used


```
def book_function(url_data):


    data=[]

    for url_2 in url_data['URL']: #[:4]: slicing only the top 4
        time.sleep(3)
        amazon = requests.get(url_2, headers=headers)
        content = amazon.content
        soup = BeautifulSoup(content)
```


```
try:
    price =(soup.find('span', attrs = {'class':'a-size-base a-color-price a-color-price'}).get_text()).strip().repl
except:
    #price = 'Nothing'
    try:
        price =(soup.find('span', attrs = {'class':'slot-price'}).get_text()).strip().replace('$', '')
    except:
        try:
            price = (soup.find('span', attrs = {'class':'a-size-base a-color-secondary'}).get_text()).strip().repla
        #price = 'Nothing'
        except:
            price = 'Nothing'
```

NEXT STEPS

CONCLUSIONS

- ➔ Create a API for easier access to everyone to run the program;
- ➔ Automatization of the API and the Web Scraping codes to retrieve the updated info from the NYTimes and Amazon Books;
- ➔ Connect with more data sources;
- ➔ Create a recommendation system based on machine learning;
- ➔ Improve Amazon web scraping;
- ➔ More detailed documentation on the GitHub.

IRON HACK

IRON
HACK

ANY QUESTION?

https://github.com/thaisternus/project-week-3-data-thieves

THANK YOU!

Data Analytics
Part-time Course