© Dr. DING Chao, 2024

MSBA7001 Exercises III

Module 1, 2024-25
HKU Business School

Contents

© Dr. DING Chao, 2024

Web Scraping

Exercise – quotes

Write a program to extract all quotes from the following page http://quotes.toscrape.com/. Print out "who says: what". For instance:

```
Albert Einstein says:
The world as we have created it is a process of our thinking. It cannot
 be changed without changing our thinking.

J.K. Rowling says:
It is our choices, Harry, that show what we truly are, far more than ou
r abilities.
```

Exercise – Titanic

From the Titanic movie page, https://www.imdb.com/title/tt0120338, extract the following info, save your result in a tuple called titanic, and print it out.

- Movie title: Titanic
- Release year: 1997
- PG-rating: IIA
- Movie length: 3h 14min

```
>>> print(titanic)
('Titanic', '1997', 'IIA', '3h 14min')
```

Exercise – movie urls

From IMDb's top rated movie page, https://www.imdb.com/chart/top, extract all 250 movie urls. An example of a valid movie url is: /title/tt2096673. Save your result in a list called movieurls. Make sure there are no duplicates in the list. Note that you would only get the first 25 movie links after IMDb's latest update.

```
>>> print(movieurls[:3])
['/title/tt0111161/', '/title/tt0068646/', '/title/tt0468569/']
```

Finally, store these urls in a text file called "urls.txt". The file should look like this:

```
/title/tt0111161
/title/tt0068646
/title/tt0071562
/title/tt0468569
```

Exercise – movie info

Open and read "urls.txt" you have created in Exercise – movie urls. For each valid url, extract the same information as you did in Exercise - Titanic. Save your result in a DataFrame called movieinfo.

	Movie name	Release year	PG-rating	Movie length
0	The Shawshank Redemption	1994	IIB	2h 22m
1	The Godfather	1972	IIB	2h 55m
2	The Dark Knight	2008	IIB	2h 32m
3	The Godfather Part II	1974	IIB	3h 22m
4	12 Angry Men	1957	I	1h 36m

Finally, store all the data in a csv file called "movieinfo.csv".

Exercise – Amazon best sellers

Build a crawler to extract book information from Amazon's top 50 best sellers at https://www.amazon.com/Best-Sellers-Kindle-Store/zgbs/digital-text

Notes:

1. Maximum rating is 5, you may see the actual rating by hovering your mouse over the stars. This value is visible in the html source code.
2. Not every book has a rating or price. For such books, keep their rating or price to be np.nan.
3. You may extract the rank from the html page or simply create it by yourself.
4. For books that have a price, extract the numeric value of price without the dollar sign ($).
5. DO NOT make too many requests to the page in a short period of time. Amazon may temporarily ban your IP address. Instead, make a soup in a separate cell, and then retrieve information from the soup in other cells.
6. It's likely that only the first 30 books can be extracted after Amazon's latest update.

Save your result in a DataFrame called bestseller.

	title	author	rating	price
0	Things We Left Behind (Knockemout Book 3)	Lucy Score	4.7	5.99
1	The Lost Bookshop: The most charming and uplif...	Evie Woods	4.5	0.99
2	The Coworker: An Addictive Psychological Thriller	Freida McFadden	4.2	3.99
3	Cruel Paradise (Oryolov Bratva Book 1)	Nicole Fox	4.7	2.99

Finally, store all the data to a csv file named "bestseller.csv".

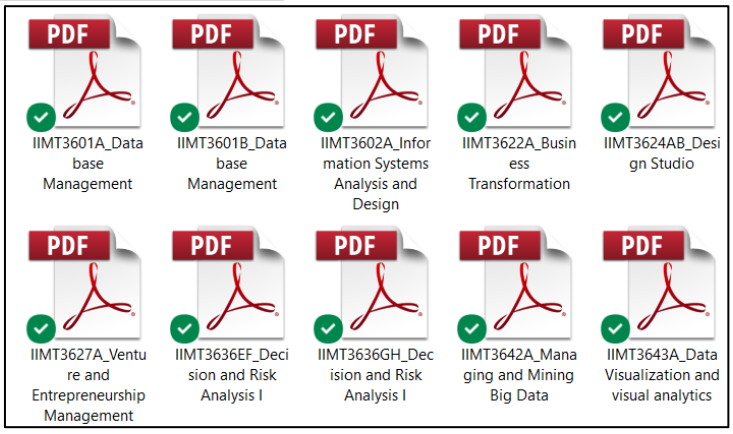
Exercise – UG course outlines

Use the course outline table from the following page: https://ug.hkubs.hku.hk/course

1. Create a DataFrame called data based on this table.
2. Replace the Outline column with the actual link to the outline document.
3. Create a sample called sample that includes only IIMT level 3 courses offered in semester 2. Make sure that the data type of the Semester column is int64.

	Code	Course Name	Lecturer	Semester	Outline
0	IIMT3601A	Database Management	Prof. Shengjun MAO	2	https://ug.hkubs.hku.hk/f/course/254218/IIMT36...
1	IIMT3601B	Database Management	Prof. Michael Chau	2	https://ug.hkubs.hku.hk/f/course/254565/IIMT36...
2	IIMT3602A	Information Systems Analysis and Design	Dr. C K LOK	2	https://ug.hkubs.hku.hk/f/course/254219/23_24-...

4. Download the outlines of courses in #3 and save them in a folder called outlines which sits inside the data_out folder. Name each downloaded outline as code_name.ext, e.g., IIMT3601A_Database Management.pdf



Hint: Python treats documents as the type of bytes. A byte consists of 8 binary numbers. Therefore, when writing a document, use the mode 'wb' (write binary). See illustration below.

```
import requests

data = requests.get(url).content
with open(filepath, 'wb') as handle:
    handle.write(data)
```

Exercise – box office mojo

Extract the latest Domestic Daily box office of the Oppenheimer movie from the following page: https://www.boxofficemojo.com/release/rl3725886209/

This is an open exercise. No requirement on what info to extract and how to save your result.

Domestic Daily | Domestic Weekend | Domestic

Date	DOW	Rank	Daily
Jul 21	Friday	2	$33,017,635
Jul 22	Saturday	2	$26,248,140
Jul 23	Sunday	2	$23,189,645
Jul 24	Monday	2	$12,671,950

Exercise – NYT headlines

Extract all the NYTime's headlines under the "News" section from the following page: https://www.nytimes.com/rss

By clicking open every region, you will see an XML file that documents the details of news headlines. This is an open exercise. No requirement on what info to extract and how to save your result.

Exercise – Paris Olympic Games

Extract the medal count from the following page: https://olympics.com/en/paris-2024/medals

This is an open exercise. No requirement on what info to extract and how to save your result. Note that you may need to use selenium.

Exercise – top 100 universities

Extract the ranking of the top 100 universities from the following page: https://www.topuniversities.com/student-info/choosing-university/worlds-top-100-universities

This is an open exercise. No requirement on what info to extract and how to save your result. Note that you may need to use selenium.

Exercise – super charge

Extract the first 5 pages of super charge information from the following page: https://supercharge.info/data

This is an open exercise. No requirement on what info to extract and how to save your result. Note that you may need to use selenium.