MSBA7001_2021-22 Assignment 2

- 5 problems, 30pts
- Keep all your data files in the same fold as your script file. Use relative path when reading and writing files.
- Save your script in a Jupyter Notebook file named "A2.ipynb"
- Compress all the files including "A2.ipynb", "A2_fitness.csv", "A2_words.csv", "A2_urls.txt", A2_books.csv", 'fitness.json", "policy.txt", and "stop_words.txt" as "A2.rar" or "A2.zip" and submit it on Moodle
- **Due Sept 18 (Saturday), 11:59pm**

Problem 1 (6pts)

Note: this was a 2019 exam problem.

As of now, there are 34 fitness walking tracks in 18 districts of Hong Kong. Each track has a track length and expected energy consumption. For instance, there is a track in the Ap Lei Chau Wind Tower Park with a track length of 1200 meters and energy consumption between 50 and 60 calories. Your job is to extract all the tracks' data and create a track map.

The "fitness.json" file has information about all the 34 fitness walking tracks. Detailed description of each track is stored as key-value pairs as follows:


```
"Title": "Ap Lei Chau Wind Tower Park",
"DistrictS": "Southern",
"DistrictL": "Hong Kong Island",
"Route": "Total Track Length: 1200 m<br>Calories consumed: 50-60 Cal",
"HowToAccess": "MTR: Lei Tung Station Exit A1<br>Bus: 90, 90B, 90C, 91, 91A, 93, 93C, 95C,
"MapURL": "https://www.lcsd.gov.hk/en/sportforall/common/graphics/en/walk/map_01.jpg",
"MapURL_tc": "https://www.lcsd.gov.hk/en/sportforall/common/graphics/b5/walk/map_01.jpg",
"MapURL_sc": "https://www.lcsd.gov.hk/en/sportforall/common/graphics/b5/walk/map_01.jpg",
"Latitude": 22.24472222,
"Longitude": 114.1525
```


length mincal maxcal

Description of the main keys is listed in the following table.

Key	Description
Title	Name of venue where the fitness walking track is located
DistrictL	Region name. Valid value(s): - Hong Kong Island - Kowloon - New Territories
DistrictS	Districts under the region (DistrictL)
Route	Track length and energy consumption
Latitude	Latitude of the track
Longitude	Longitude of the track

From “Route”, one may find the track length (as length), minimum calories (as mincal), and maximum calories (as maxcal) required on the track. Read the json file and extract all related information of each track. Write the data to a csv file named “A2_fitness.csv”. Your csv file should look like this when opened in Excel.

title	districts	districtl	latitude	longitude	length	mincal	maxcal
Ap Lei Chau \	Southern	Hong Kong Is	22.2447222	114.1525	1200	50	60
Sun Yat Sen	Central and Western	Hong Kong Is	22.290402	114.143783	2430	95	115
Hong Kong P	Central and Western	Hong Kong Is	22.27714	114.163813	1200	45	50
Wan Chai Pa	Wan Chai	Hong Kong Is	22.275472	114.176008	321	10	15

It has 35 rows (including the header). The header should be: “title, districts, districtl, latitude, longitude, length, mincal, maxcal”. **Requirement: use the csv module to write to the csv file.**

Problem 2 (6pts)

Note: this problem is modified from a 2018 exam problem.

On Oct 10th 2018, the Chief Executive of the HKSAR, Mrs. Carrie Lam, gave a policy address to the Legislative Council. The policy address lays out future plans for housing, health care and more. The speech is available at: https://www.policyaddress.gov.hk/2018/eng/speech.html

The entire speech is saved in the text file "policy.txt". Your goal is to find the frequency of each unique word in the text.

To obtain clean data, you need to remove punctuations before making the count. The most common English punctuations can be found as follows:

```
import string
string.punctuation
```

The output is a string that looks like this:

```
!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~
```

In the speech, there are also Chinese punctuations as follows:

```
“ ”  ‘ ’
```

You need to remove all such English and Chinese punctuations from the text. To obtain meaningful data, you also need to remove stop words. Stop words refer to the most common words in a language such as "the," "from," and "are" in English. These words appear to be of little value, thus need to be removed. Read the text file "stop_words.txt", where it has a list of common English stop words. Remove all the stop words from your answer. For simplicity, you do not need to handle any other punctuations or variations of words such as "that's" vs "that", "yours" vs "your".

Count the frequency of each unique word. Save your answer as a csv file named "A2_words.csv". The header should be "word, frequency". In Jupyter Notebook, show all the words with a frequency higher than 20 as the following graph, which may also help you verify your own answer. **Requirement: use pandas' DataFrame to write to the csv file.**

	word	frequency
6	POLICY	32
7	ADDRESS	22
25	HONG	50
26	KONG	44
49	HOUSING	35
50	LAND	22
129	DEVELOPMENT	27
153	NEW	23
176	PUBLIC	23
211	GOVERNMENT	45

Problem 3 (6pts)

Write a program to read the page: https://www.imdb.com/chart/top. Extract all the valid links and user ratings.

A valid link looks like: /title/tt3170832/
A user rating looks like: 9.2 based on 1,362,495 user ratings

Save all the matched valid links in a text file named "A2_urls.txt". Make sure that there are no duplicates in the file. In total, you should have 250 unique links. Your file should look like the left figure, but may not be exactly the same. You are recommended to clear your browser's browsing history before scraping. Store all the matched user ratings in a list called user_ratings. In Jupyter notebook, print the number of unique movies and the first 6 movies' ratings as output (right figure). **Requirement: use only regular expressions for data extraction (do not use BeautifulSoup 4). Use loops, instead of built-in functions, to remove duplicates.**

```
/title/tt0111161/
/title/tt0068646/
/title/tt0071562/
/title/tt0468569/
/title/tt0050083/
/title/tt0108052/
/title/tt0167260/
```

Left

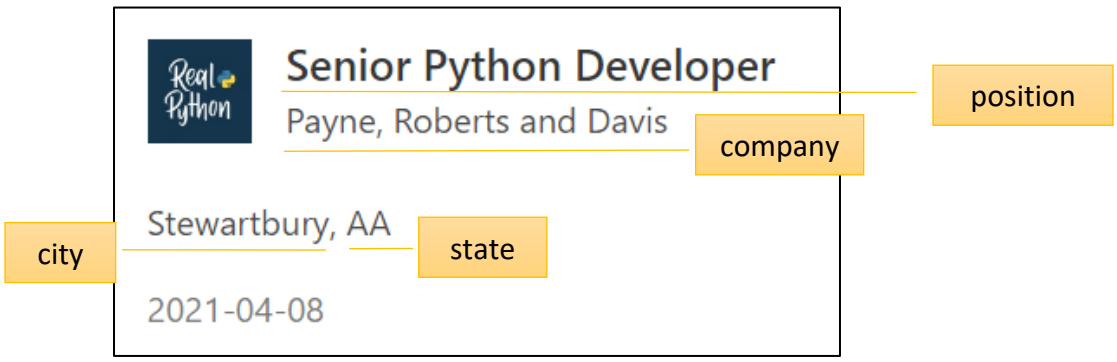
```
250 unique movies in the top chart.

Here are top six movies' user ratings:
No. 1: 9.2 based on 2,454,954 user ratings
No. 2: 9.1 based on 1,698,828 user ratings
No. 3: 9.0 based on 1,179,886 user ratings
No. 4: 9.0 based on 2,410,657 user ratings
No. 5: 8.9 based on 726,757 user ratings
No. 6: 8.9 based on 1,262,600 user ratings
```

Right

Problem 4 (6pts)

Build a crawler to extract the fake jobs information at: https://realpython.github.io/fake-jobs/
Copy the link to your browser if it can't be clicked open.

Use BeautifulSoup 4 to retrieve the following information:



Create a DataFrame to store your data. There should be 100 jobs. In Jupyter Notebook, show all 4 "engineer" related jobs in "AA" state. Your output should look like this:

	position	comany	city	state
1	Energy engineer	Vasquez-Davidson	Christopherville	AA
28	Structural engineer	Pierce-Long	Herbertside	AA
32	Broadcast engineer	Morgan, Butler and Bennett	Loribury	AA
48	Engineer, broadcasting (operations)	Taylor PLC	Gileston	AA

Problem 5 (6pts)

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. Some books' title may include series number as indicated by a pair of parentheses (e.g., The Miles High Club Book 3). When extracting title, exclude such information. The title should simply be "The Casanova".
2. 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.
3. Not every book has a rating. For such books, keep their rating to be NumPy's "not a number".
4. You may extract the rank from the html page or simply create it by yourself.
5. Extract the numeric value of price. Do not include the dollar sign ($).
6. 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. If urlopen(url) does not work, then try requests.get(url).content.

In Jupyter notebook, show all 5 books whose rating is higher than 4.5 and whose price is exactly 4.99.

	rank	title	author	rating	price
0	1	The Keeper of Happy Ending	Barbara Davis	4.6	4.99
2	3	Timber	Tate James	5.0	4.99
9	10	The Highland Flin	Meghan Quinn	4.6	4.99
31	32	Blood & Bones: Ozzy	Jeanne St. James	4.9	4.99
44	45	Drilled	K.M. Neuhold	4.8	4.99

Write your data to a csv file named "A2_books.csv". You should retrieve 51 rows in your csv file, including the header. You may use either the csv module or pandas' DataFrame to write the data. **Requirement: use only CSS selector for tag selection.**