

ITI102 Assignment 2 (Total 40 Marks)

Instruction

- 1. This is an individual assignment.
- 2. The solution must be implemented using Python 3 codes with the Colab notebook.
- 3. Answer all the following questions.
- 4. You must zip up the codes into a single zip file for submission in BrightSpace.
- 5. Submit your answers by 2 July 2023 23:59hr

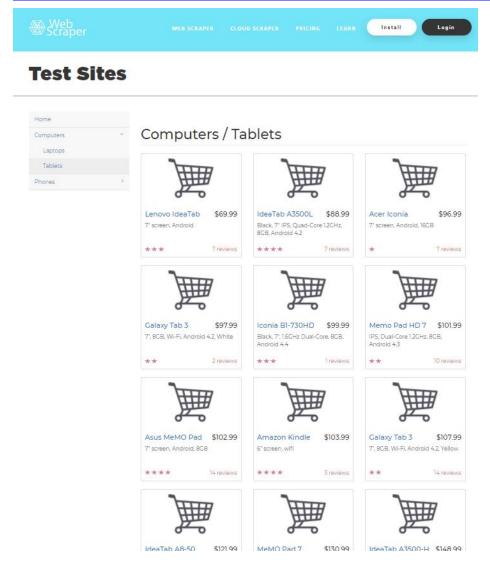


Questions

Question 1 (6 marks)

a) Use the *python Scrapy* module to implement a web scraping program to get the content from the following websites (4 marks)

https://webscraper.io/test-sites/e-commerce/allinone/computers/tablets

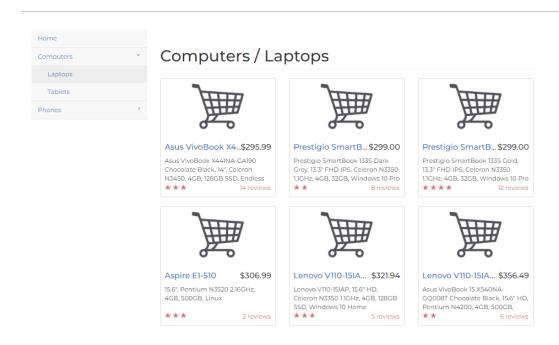




https://webscraper.io/test-sites/e-commerce/allinone/computers/laptops



Test Sites



The URL above points to an e-Commerce store that sells different tablet/laptop models. The purpose of the site is for web scraping test.

You must collect information for all products listed on the webpage.

Each product should storage with the information: product type, description, price and review information.

The result of the scraped data must be stored in a JSON format file. An example is as follows:

{"type": "tablet", "price": ["\$603.99"], "description": ["Wi-Fi, 64GB, Silver"], "product": "Apple iPad Air", "review": "7 reviews"}

{{"type": "laptop", "price": ["\$1272.99"], "description": ["Silver, 12" IPS, Quad-Core 2.2Ghz, 16GB, 4G, Window 10"], "product": "IdeaTab S5000", "review": "8 reviews"}

b) Develop a python function to search for a product information based on the review. (2 marks)

Print your function result with review=8 and review=14.



Question 2 (9 marks)

Implement a Singapore traffic report system using python.

The program must be able to get data from the Singapore LTA data link as shown below.

Read the road incidents data from the following API(Application Programming Interface)

http://datamall2.mytransport.sg/ltaodataservice/TrafficIncidents

Read the road traffic bands data from following the API

http://datamall2.mytransport.sg/ltaodataservice/TrafficSpeedBandsv2

Display the data in a visualization graph.

The graph should display the Singapore map with different markers that indicate the traffic incident and traffic bands at each location.

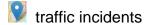
Complete the following tasks:

a)Python program request for the traffic incident using URL and get the return JSON data. Extract the traffic incident data and display in the Singapore map (3 marks)

b)Python program request for the traffic band using URL and get the return JSON data. Extract the traffic band data and display in the Singapore map (3 marks)

c)Use different markers to represent and differentiate the traffic incident and traffic bands in the map. Allow the markers to display more information when it required(3 marks)

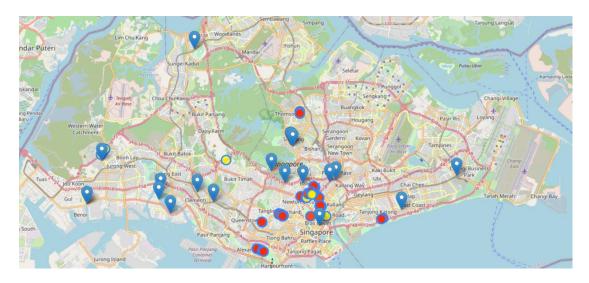
The following example shows an example of a visualization map with data markers.



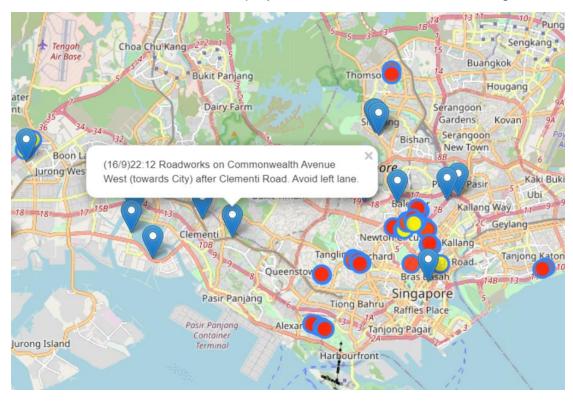
and different speed band(you may use additional makers for more speed bands)

The additional markers should be distinct from the originals.





If a user clicks on the marker, display relevant information in the dialog box.

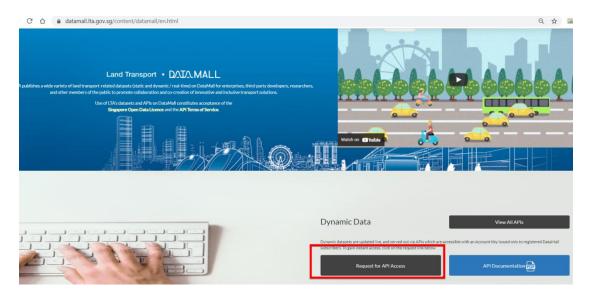


Resources

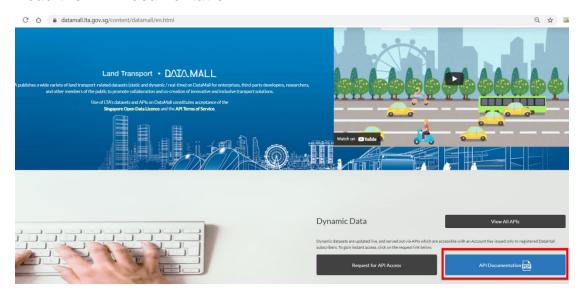
This is the LTA data provider link and documentation https://datamall.lta.gov.sg/content/datamall/en.html

Register for an API





Read the API Documentation



First API is to get traffic incidents (page 34)





LTA DataMall | API User Guide & Documentation Version 5.2 (28 May 2020)

2.19 TRAFFIC INCIDENTS

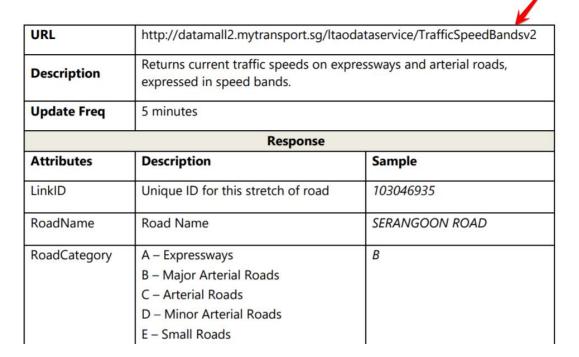
URL	http://datamall2.mytransport.sg/ltaodataservice/TrafficIncidents				
Description	Returns incidents <u>currently</u> happening on the roads, such as Accidents, Vehicle Breakdowns, Road Blocks, Traffic Diversions etc. 2 minutes – whenever there are updates				
Update Freq					
	Response	E			
Attributes	Description	Sample			
Туре	Incident Types:	Vehicle Breakdown			

The second API is TrafficSpeedBandsv2 (page 35)



LTA DataMall | API User Guide & Documentation Version 5.2 (28 May 2020)

2.20 TRAFFIC SPEED BANDS





Question 3 (9 Marks)

You are given a set of text data that expresses the sentiments of customers. The sentiments are label as follow:

pos-positive

neg- negative

The text data are stored in the Train and Test folders with two subfolders pos and neg. In each of these, the subfolder contains 100 text reviews.

Refer to the Q3sentimentClassificationV2_student.ipynb.

Complete the data preprocessing tasks in the ipynb file using python Natural Language Toolkit(https://www.nltk.org/).

Question 4(10 Marks)

Given the following dataset1.csv

Here show the first 5 rows of a population data

Unnamed: 0 age educatn earnings hours kids married

0	0 39	12.0	77250 29	940 2	married
1	1 35	12.0	12000 20	040 2	divorced
2	2 33	12.0	8000 69	93 1	married
3	3 39	10.0	15000 19	904 2	married
4	4 47	9.0	6500 168	33 5	married

Given the follow Hypothesis

Null Hypothesis HO:

Work hours for people with higher earnings == Work hours for people with lower earnings

Alternative Hypothesis HA:

Work hours for people with higher earnings > Work hours for people with lower earnings

Conduct the hypothesis test with sample data using python scipy.stats function.

State your result of the hypothesis test.



Question 5(6 marks)

An IOT device collected data in the following format dev-144 temperature is 40.5 degree celsius, pressure is 20000.5 pa dev-146 pressure is 24000.5 pa, temperature is 35.5 degree Celsius dev-23 temperature is 33.5 degree celsius, pressure is 10400.3 pa dev-25 pressure is 40040.5 pa, temperature is 24.5 degree Celsius

Use python regular expression to extract the data and store into a Json format.

{ "device": "dev-144", "temperature": 40.5, "pressure": 20000.5}

End of questions