ITI102 Assignment 2 (Total 40 Marks)

Instruction

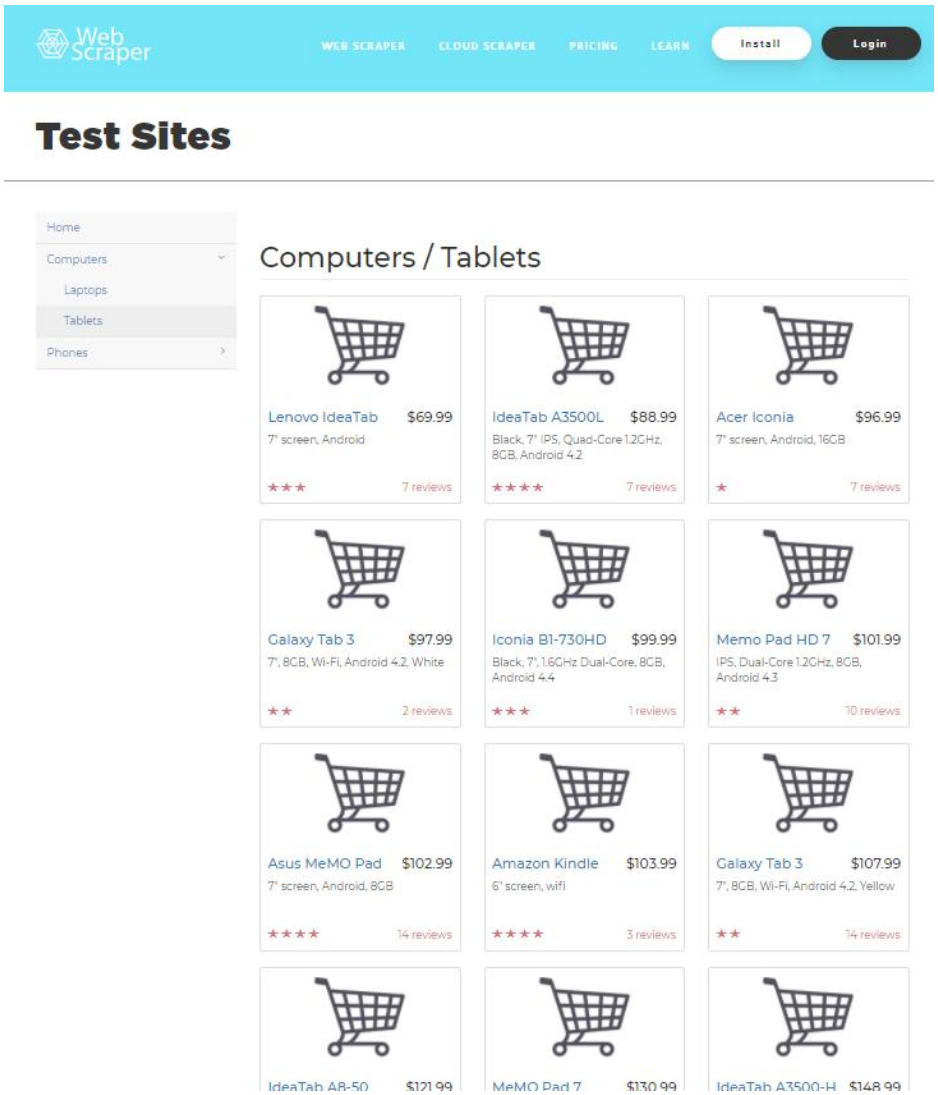
1. This is an individual assignment.
2. The solution must be implemented using Python 3 codes with the Colab notebook.
3. Answer all the following questions
4. You must zip up the codes into a single zip file for submission in BrightSpace.
5. Submit you answers by 3 July 2022 23:59hr

Questions

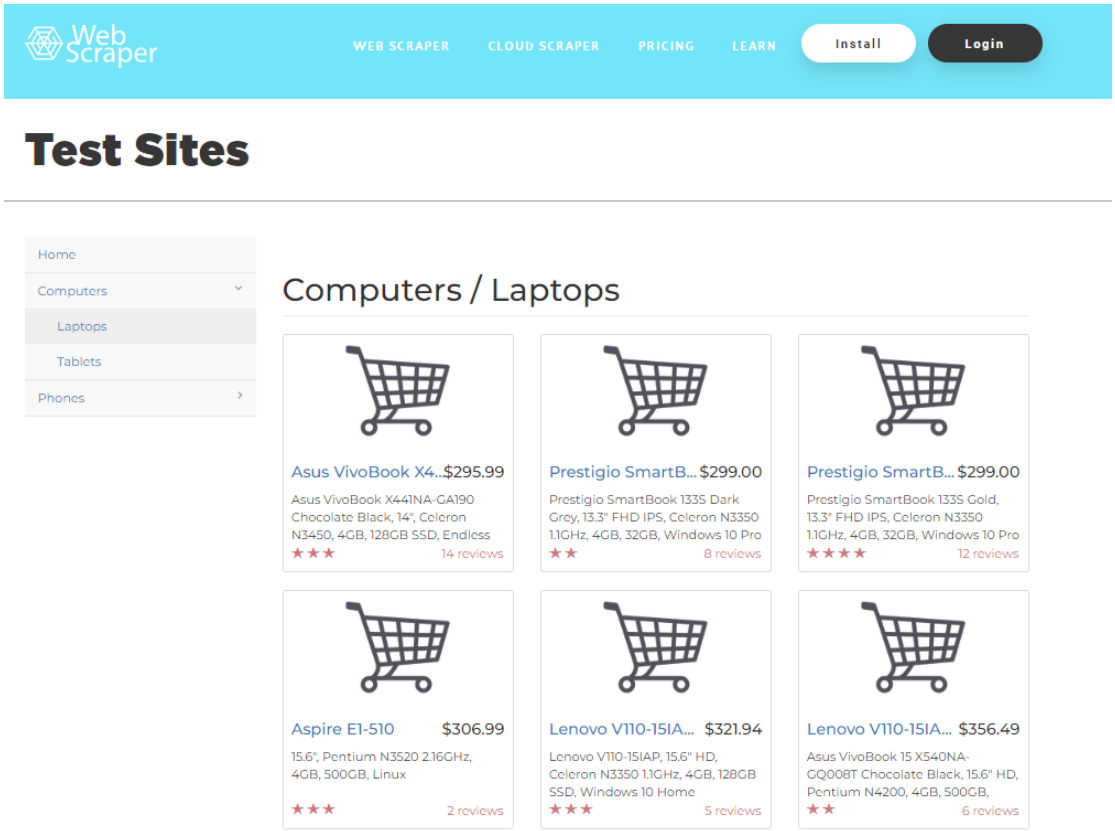
Question 1 (9 marks)

a) You must use the ***python Scrapy*** module to design a web scraping program to get the content from the following websites (6 marks)

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



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



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

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

You are required to collect product, description, price and review information of all the tablets listed.

Use python scrapy framework in your python program.

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 tablets’ information based on the review. (3 marks)

Function name: SearchbyReview(int review)

Argument review: int

Return result: list of all matching items with (type, product, description, price, reviews) that have review greater than or equal to the function argument review. The list needs to be sorted base on the reviews in the descending order.

Run your function with review=8 and review=14. Print the results of each of the review.

Question 2 (12 marks)

Design a Singapore traffic report system using python.

The program must be able to collect 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 collected 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.

Marking criteria

1. Python program request for the traffic incident using URL and get the return JSON data (2 marks)

2. Extract and format the JSON traffic incident data to be displayed in the Singapore map (2 marks)

3. Python program request for the traffic band using URL and get the return JSON data (2 marks)

4. Extract and format the JSON traffic band data required for displaying in the Singapore map (2 marks)

5. Add the formatted data in the map using different markers to represent the traffic incident and traffic bands (2 marks)

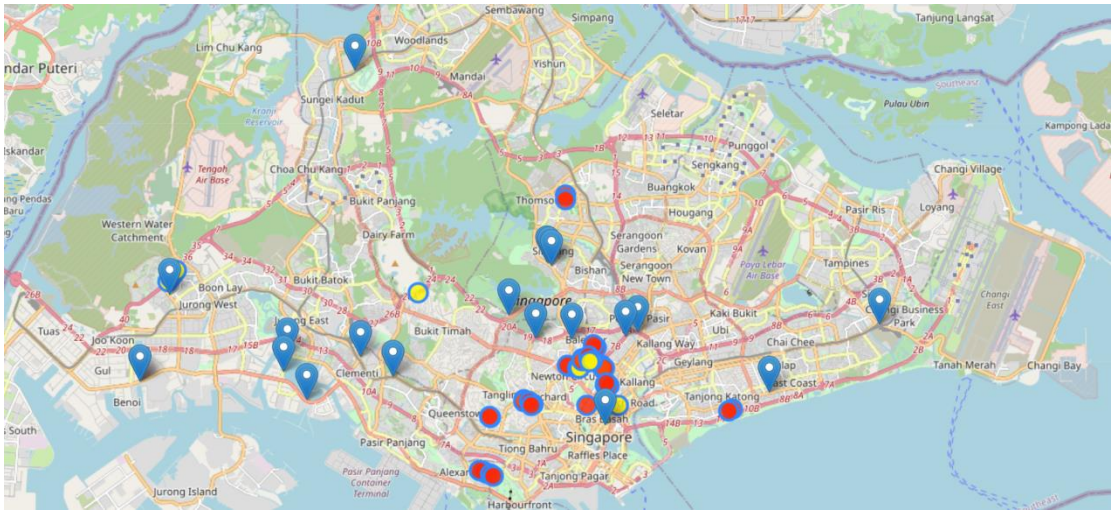
6. Display the traffic incident or traffic band information when the marker is clicked(2 marks)

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

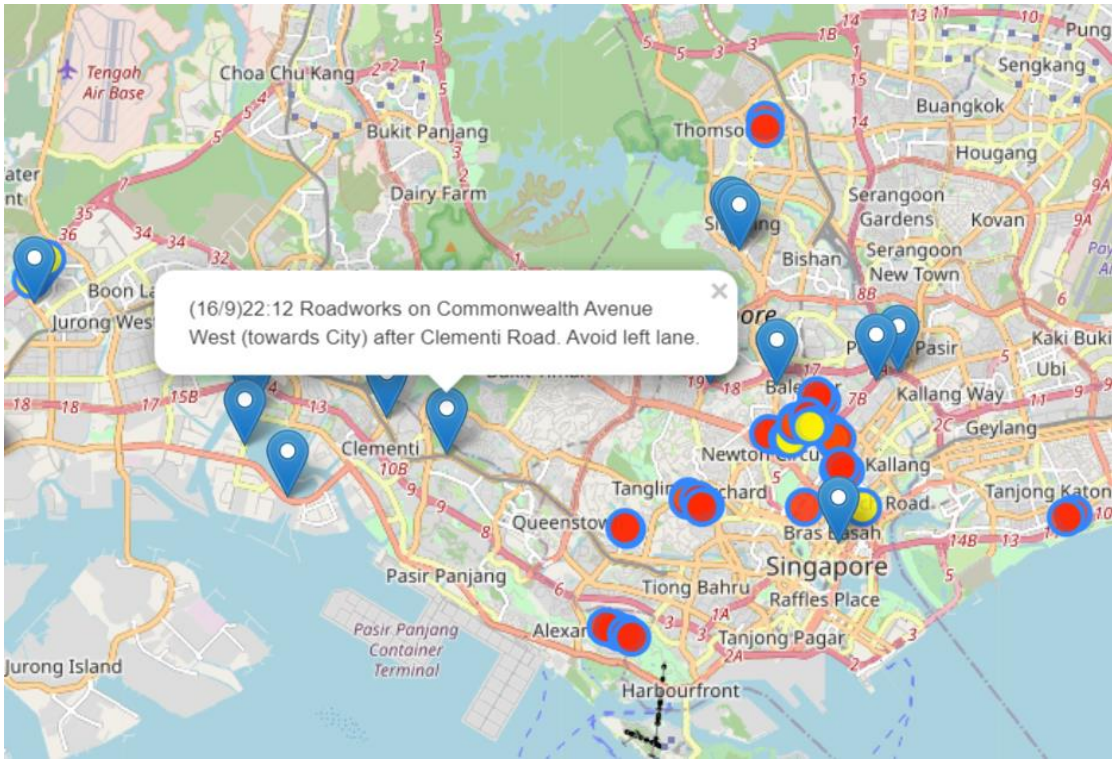
traffic incidents

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.

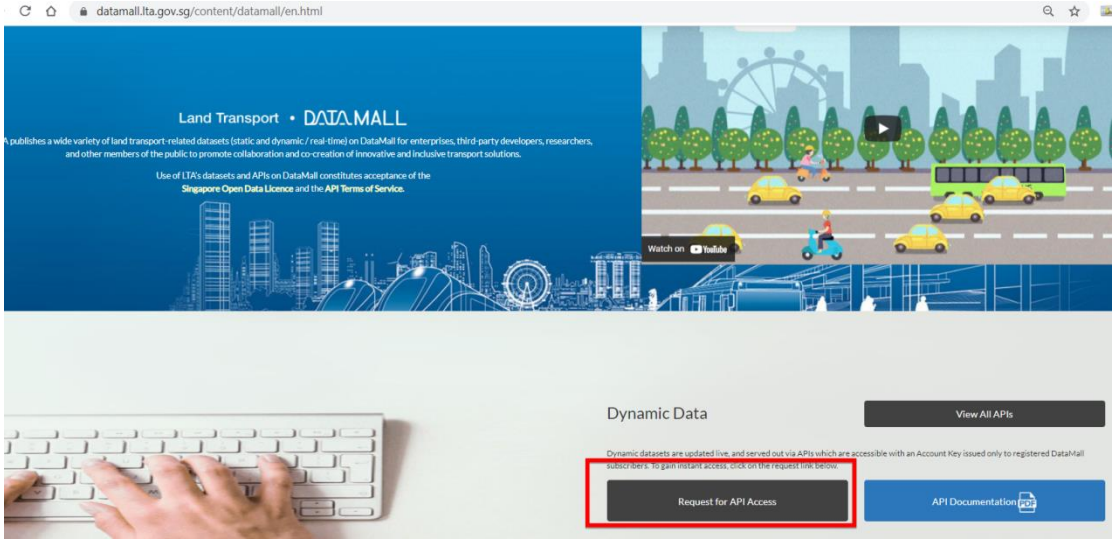


Resources

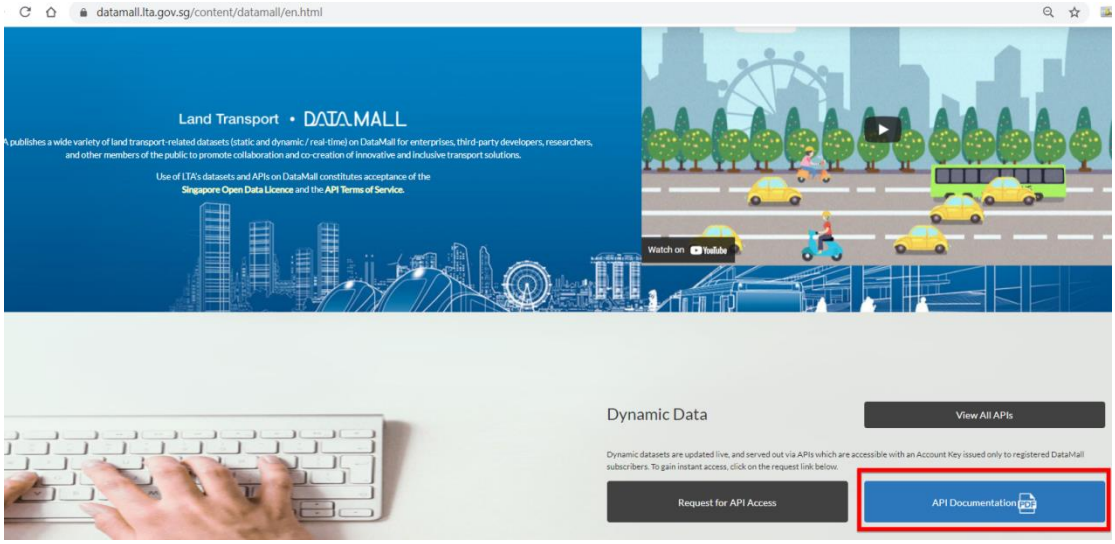
This is the LTA data provider link and documentation

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 currently happening on the roads, such as Accidents, Vehicle Breakdowns, Road Blocks, Traffic Diversions etc.	
Update Freq	2 minutes – whenever there are updates	
Response		
Attributes	**Description**	**Sample**
Type	Incident Types: • Accident • Road Works • Vehicle Breakdown • Weather • Obstacle • Road Block • Heavy Traffic • Misc. • Diversion	*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

URL	http://datamall2.mytransport.sg/ltaodataservice/TrafficSpeedBandsv2	
Description	Returns current traffic speeds on expressways and arterial roads, expressed in speed bands.	
Update Freq	5 minutes	
Response		
Attributes	**Description**	**Sample**
LinkID	Unique ID for this stretch of road	*103046935*
RoadName	Road Name	*SERANGOON ROAD*
RoadCategory	A – Expressways B – Major Arterial Roads C – Arterial Roads D – Minor Arterial Roads E – Small Roads	*B*

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 Q3sentimentClassification_Question.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	2940	2	married
1	1	35	12.0	12000	2040	2	divorced
2	2	33	12.0	8000	693	1	married
3	3	39	10.0	15000	1904	2	married
4	4	47	9.0	6500	1683	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 you result of the hypothesis test.

----------End of questions----------