INDIAN SCHOOL OF BUSINESS

Data Collection & Pre-Processing

Individual Assignment

Instructor: Dr. Manish Gupta

Honor Code Scheme: **2N-b** Weightage: **40%**

Deadline**: 25th March 2023,11:55 PM**

Deliverables:

1. Submit either .py format code files **or** ipynb format code files.
    - If submitting **.py format** code files, then **also submit a pdf document for the report** which **includes the output** for each part of the question in the report. (Give proper comments/explanations/interpretation for the code)
    - If submitting Python notebook **(.ipynb format**) code files then **retain the outputs in the notebooks** and give proper comments/explanations/interpretation for the code)
2. If you have multiple .py or .ipynb files, or if your .py file takes command line arguments then submit a **README file** with steps to run the pieces of code.
3. Two result files: **tableList.tsv**, **tableDetails.tsv**
4. Assignment Submission Form

General Instructions:

1. This is an Individual Assignment.
2. **Do NOT** submit **.zip** files otherwise the submission will not be considered.
3. Any **late submission will attract a penalty** as mentioned in the course outline
4. The honor code for this submission is **2N-b**. **Please look through the honor code restrictions carefully before attempting the assignment as there will be strong consequences for breaking them.**
5. **Email submissions are not allowed. All the submissions must be made on the LMS.**
6. Upload your submissions to the '**DCPP Individual Assignment'** folder **on LMS**.
7. **Please adhere to the given instructions, otherwise, your submission will not be accepted, or a severe penalty will be applied.**

Collecting Product Data from eBay

Goal: To understand basic web page scraping.

Expected time: 7-8 hours.

Requirements: You need to have a machine with Internet connectivity. You should write code in Python. You are not allowed to use automated scraping tools like Octoparse.

This is an individual assignment.

The task is to collect data from ebay.com. The assignment has 2 parts: (1) collect information about a list of tables from eBay (2) collect detailed information about each table. Wherever there are multiple things involved per column, use a comma delimiter.

Part 1:

From this page: https://www.ebay.com/sch/i.html?_nkw=table, scrap the following information

1. Product name
2. Price
3. Shipping cost
4. Image URL
5. Number of watchers
6. Product condition (Brand new, Opened Box or Pre-owned)

Save results in a file tableList.tsv

Part 2:

For each product, go to the product description link like https://www.ebay.com/itm/Chess-Table-4-in1-Handmade-Decorative-Wooden-Chess-Backgammon-Poker-and-Checkers/184035173371 and extract the following:

1. Product Name
2. Return Policy
3. Ships to
4. Item location
5. Shipping cost
6. Estimated delivery date
7. Payment modes available
8. Price
9. Starting bid
10. eBay item number
11. Condition
12. Brand
13. Color

14. Type
15. Material

Save results in a file tableDetails.tsv

Other notes:

1. Selector Gadget tool will not get you everything. You will have to try out various things from what you learned in class.
2. You can use **Beautiful Soup** or **Selenium** to scrap content from web pages.
3. You will need to download relevant pages.

Due Date: 25th March 2023,11:55 PM