# AIML Feature ENGINEERING Projet 1

Submission Date: 03 June 2020 11.59 PM

Weightage: 30%

Imagine you have to pull a large amount of data from websites and you want to do it as quickly as possible. How would you do it without manually going to each website and getting the data? Well, “Web Scraping” is the answer. Web Scraping just makes this job easier and faster.

Web scraping is a term used to describe the use of a program or algorithm to extract and process large amounts of data from the web. Whether you are a data scientist, engineer, or anybody who analyzes large amounts of datasets, the ability to scrape data from the web is a useful skill to have. Let's say you find data from the web, and there is no direct way to download it, web scraping using Python is a skill you can use to extract the data into a useful form that can be imported.

As a part of this project, you will be preparing your own datasets about the various daily essential (or may be fast moving consumer good FMCG) products sold on various ecommerce platforms like (but not limited to) bigbasket, grofers, reliance smart, dmart, amazon pantry etc. The main objective of this project is to give you real life experience while doing data acquistion, data integration, data cleaning and data transforamtion before attempting any of the analytical activity on the data.

the four tasks that you will be doing as a part of this exercise will be as follows:

1. Data acquisition

* You as team of three or four students has to acquire the data from three or four ecommerce platforms using the technique of web scrapping.
* further on you can concentrate your search based on some of the categories of products like vegetables/fruits, grocery items, branded items etc.
* outcome of this step will be three or four different data sets each one belonging to one platform.
* document all your efforts appropriately in the jupyter notebooks with description and code.
* the weigtage for this task will be 15 marks

1. data cleaning

* the three or four datasets you have gathered might have some quality issues in them like (but not limited to) missing values, duplicate records, AND DERIVED attributes.
* you have to cleanse your datasets to remove all such daunting isssues.
* narrate all the issues which you encounter during this exercise clearly with appropriate explanation and code.
* the weigtage for this task will be 5 marks.

1. Data integration

* by this time you must have cleansed data available with you. for further processing you need to integrate all these datasets together into a single dataset.
* you may need to decide upon a common schema for this activity which can be applied on the all datasets. you may think of adding or deleting or modifying the attributes of the existing datasets.
* capture all the thinking gone behind preparing such a single database in the descriptive manner in the jupyter notebook along with the code.
* the weigtage for this task will be 5 marks.

1. exploratory data analysis and recommendation

* use the exploratory data analysis technique on the dataset in order to find out the interesting insights that are hidden within the data captured.
* now assume that a customer is looking for a particular item, then with help of simple program you should be able to recommend the online platform from which he should make a purchase of this item.
* describe all eda steps those are done with the observations obtained out of it with the help of python code in jupyter notebook.
* the weigtage for this task will be 5 marks.

Notes:

* This is a take-home PRoject to be carried out by group OF LEARNER.
* as per the need, the demos / vivas can be arranged further on.
* you are free to make selection of appropriate platforms for this exercise as long as its related to the daily essential (or fmcg) products.
* wherever required you can make appropriate assumption but make sure that you have spelt them appropriately in the submitted documents.
* This is programming exercise - requiring the approach of web scrapping with python to be used for coleecting the required data.
* You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.
* You have to write the appropriate Python code in Jupyter notebook to support you answers and submit with following nomenclature

- FE\_project1\_<group\_ID>.ipynb

* In case of any further queries, if those are generic once, learners are encouraged to use discussion forums, otherwise they can reach out to me at [ppawar@wilp.bits-pilani.ac.in](mailto:ppawar@wilp.bits-pilani.ac.in).
* Manage your efforts properly as there is no scope to shift the deadlines announced above.

References:

1. [Real Python Web Scrapping tutorial](https://realpython.com/beautiful-soup-web-scraper-python/)
2. [DataCamp Web Scrapping tutorial](https://www.datacamp.com/community/tutorials/web-scraping-using-python)