DCPP Group Assignment

This component has 50% weightage in the course

**Problem:** Create a dataset using an end-to-end data collection and pre-processing pipeline for a specific domain.

**Input:**

1. Name of the domain/subdomain for which you need to collect data and process the data into a structured source.
2. A seed set of structured sources (optional)
3. A seed set of unstructured sources (optional)

Example domains: Information about all BSE/NSE listed companies, NASDAQ companies, All publicly listed companies in energy sector, All the MPs/MLAs/MLCs in state and central government, all academy, golden globe award winning movies, all books that won prizes, all types of fish, all stars, all car models, all flight models, etc.

**Output:**

A structured data source of required information in that domain. But the storage can be a Jason file that can be read/written by Panda Data frame within Python

**Steps involved (detailed guidelines in the next section):**

1. Understand the domain, seed set of structured and unstructured sources
2. Add more structured and unstructured sources that are available in open domain
3. Download/crawl/collect data from all the sources
4. Convert data from original sources (Webpages, pdf files, CSV files, …) to structured data fields
5. Data cleaning/pre-processing as needed
6. Strategy to enhance the data with crowd sourcing methods

**Note:**

1-5 involved actual implementation

6 involves thinking /strategy/design (ready for implementation).

**Submission:**

1. Dataset in JSON format
2. Report
3. GitHub link for all your data/source code/other documentation.

**Data Collection Guidelines:**

To gather domain-specific data and the most critical objective of data collection is ensuring that information-rich and reliable data is collected. The approach of data collection might not be the same for all domains, it would mostly depend on the type of information/ data available for that domain. For example, some domains might have easily available data sources like spreadsheets, csv files, etc., while others would require scraping from different sources to form a reliable database. Irrespective of that, you will now see a generic approach that you could incorporate while collecting data.

* You need to first estimate the number of attributes (columns) you would like to have in your data collection. This number should be reasonable – not to high not too low. For example, for a domain like movies, the attributes can be the year of release, title, country, language, budget, production company, plot etc. Having a range of 15-50 attributes will be a good start.

**Note:** Attribute count can vary based on the selected domain and the data.

* Start searching for data sources. These can be websites, API’s, Spreadsheets, CSV files, PDF’s, Wiki dumps, Databases, etc. When a data source is found,
  + Check whether the data is meaningful and can be used for article generation. If you find it meaningful, then check for the attributes that can be used.
  + Extract the data.
    - Scraping libraries/ tools: Beautifulsoup, scrapy, tabula, selenium etc. These are some of the starter libraries/ tools you can make use of for scraping
    - **Note:** Apart from the above-mentioned libraries/ tools, there are many other ways to perform this task. Feel free to explore and make use of them.
* You can also rely on using Wikipedia articles, Wiki data for scraping the data that you want to make use of in your dataset.
* In case of API’s, Spreadsheets, CSV files we can use the data directly after analyzing it
* Images can also be collected if required for the data.
* The above steps of searching for data sources, extracting data would repeat until you have enough attributes
* If the data is collected from multiple sources, then to create a unified knowledge base we merge all the data collected (based on a primary key like the ID, name, etc).
* The data can be stored as json, excel sheets or any of the key-value data types.
* After collecting the data, you need to analyze, clean the data before making use of it.
  + You need to look for missing values, duplicate entries in the dataset. You must make sure that there are no duplicate entries. For handling the missing values, you can either remove the records with missing rows or fill the missing values depending on the data.
  + To perform Exploratory Data Analysis (EDA), we can make use of libraries like Sweetviz (open-source python library). You can refer to that [here](https://pypi.org/project/sweetviz/).

---------

**Division of Groups & Selection of Domain**

1. All the details of your group members are available in the .xlsx attached.
2. The selection of domains can be done by viewing the <https://elearn.isb.edu/mod/url/view.php?id=145324>.
3. You're free to choose another domain that is relevant to you. However, in that case, the seed data would not be available.
4. Please note, all ties will be broken based on email timestamp and preference orders. That makes this essentially first come-first serve. To not miss out on a domain you prefer, please make sure you coordinate early on and submit your preferences as soon as you can.

**General Instructions:**

1. This is a group assignment.
2. Do NOT submit .zip files otherwise the submission will not be considered.
3. Please include your team member names and PGIDs in the submission.
4. One person should submit on behalf of all the team members.
5. Late submission is applicable as per the course outline.
6. The honour code for this submission is **2N-b**.

**Deadline: 29th May 2022, 11:55 pm.**