**PubMed Data Pipeline Assignment**

A simple documentation for PubMed project

**Problem Statement**

A company is interested in analyzing the latest research on a specific medical condition and wants to gather information on the authors of relevant articles in the PubMed database. The goal is to create a data pipeline that will extract this information, clean and transform it, and then load it into a tabular format for further analysis.

1. Extract articles related to a specific medical condition - "Ulcerative Colitis".

2. Identify and extract all of the authors' relevant information for each article.

3. Load the cleaned and transformed data into a tabular format (e.g., a CSV file). The format should be in columns: firstName, lastName, affiliation (Example: John, Doe, Massachusetts Institute of Technology, Cambridge, MA, USA), etc....

4. Compare the data against a provided database of individuals. Add a boolean column indicating whether each person matches an entry in the provided database, confirming the author's existence in our records.

**Steps Taken**

**1. Accessing PubMed API**

Accessing PubMed API to extract articles and author information related to medical condition - "Ulccretive Colitis".

**2. Author Information extracting from articels**

For each recieved article from PubMed API author information is extracted including first name, last name and affiliation.

**3. Data cleaing, Transformation and structuring**

Duplicate rows were removed, missing vallues were filled by "Unknown". The cleaned data was structured in tabular format and it is including the columns 'firstName', 'lastName' and 'affiliation' then the data was exported to CSV file.

**5. Data Matching**

The transformed data was compared with provided database 'person\_profiles.json'. I found that there is a column named formated name in 'person\_profiles.json' i used that column to compare both files if there are any matches but i took some several steps before that. First i read 'person\_profiles.json' into pandas dataframe, then i checked for duplicates and missing values and then i saved the dataframe to CSV file.

For the actual matching i combined columns 'firstName' and 'lastName from the data that i extract from PubMed API and i checked if there any matches in column 'formattedName' from 'person\_profiles.json' i combined the result to a new datframe and i add column match(If there is a match the value of the column will be True else False). I saved the result to new CSV file inlcluding 'firstName', 'lastName', 'affiliation', 'fullName' and 'match'.

**Sample Data:**



**Challenges and solutions:**

**Matching:** Matching authors from extracted data with persons\_profiles.json - combining first and last name from extracted data and using column formattedName from person\_profiles.json.

**API Rate Limits:**

**API returning XML:** For that challenge i used **xmltodick** library. It is parsing to python dictonary the recived xml from the API