DCPP Group Project Report – Group 5

Rohit Sanamvenkata (12110046)
Manjari Shrivastava (12110082)
Subham Kundu (121100108)
Chaitanya Dadhich (12110022)

Executive Summary

a. Problem Statement:

A structured data source as a JSON file of chosen domain for which data needs to be collected and pre-processed into a structured source using seed set of structured/unstructured sources.

b. Brief understanding of challenges:

- Selection and understanding domain
- Data collection using structured and unstructured sources in open domain by identifying relevant attributes from each source.
- Iterative process of extracting the data from different sources.
- Merging, cleaning, and pre-processing the data.
- Doing Exploratory Data Analysis (EDA) on cleaned data set if any.
- Creating strategy for crowdsourcing of data.

c. Proposed Solution:

- Explore seed sources and attempt to crawl for additional URLs or scrape data directly
- Scrape data from URLs obtained from crawling
- Identify other relevant sources of data (similar to seed source)
- Collate data from found sources, append overlapping data to minimize duplicate rows
- Define a strategy to supplement or validate data by crowdsourcing information from surveys of hospitals, therapy area associations, foundations, and other sources

Chosen domain and Seed sources:

We chose **DISEASE** domain, with following reasons:

1. There are a wide variety of diseases with distinct characteristics, affecting varying types of populations across the world
2. Health and disease information is published by various governmental, academic and research organizations – however, this data is unstructured and disorganized

3. The objective behind gathering disease data was to create a structured data set across the spectrum of disease types – infectious, lifestyle, genetic/hereditary diseases, metabolic, etc.

Seed sources and Analysis conducted:

1. We used Malacards as our starting point, however, we faced challenges getting the type of data we were interested in
2. We explored other sources with the objective of finding disease characteristics in terms of qualitative and quantitative data
3. We analysed additional sources such as Kegg, Mayo clinic, Wikipedia, Kaggle data sets, etc. in terms of their information richness and volume – we captured data on symptoms, severity, chromosomes/genes involved, infectious agent, transmission methods, prevalence, availability of vaccines or drugs, and other relevant attributes

Structured and unstructured sources from open domain/ internal sources:

What are the sources chosen and explain the reasoning behind the choices.

We have chosen below structured and unstructured sources from open domain:

a. **Structured Sources:**
    - **Wikipedia:** Being one of the most famous crowdsourced databases, the data available here is always updated and validated by multiple users across the globe. Hence, it contains credible data.
    - **Public Health Agency – North Ireland**: Since, this website is maintained by North Ireland's government, this has sanitized and credible data. Along with it, it also helps us give a list of diseases which may be more prevalent in North Ireland giving us an edge of geographical considerations.
    - **HealthEd** – **New Zealand:** This website is maintained and updated by Govt. of New Zealand and gives us disease prevalent in Asia region.
    - **Kaggle**: Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Lot of crowdsourced data is available here with people doing lot of EDA on datasets.

b. **Unstructured Sources:**
    - **Mayo Clinic:** Being one of non-profit American academic medical centre focused on integrated health care, education, and research, established in 1889, this is one of most established sources of information regarding diseases highlighting its causes, symptoms and prevention.
    - **KEGG:** KEGG database is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. Scraping data from KEGG helped us to get attributes related to genomes, drugs, and carcinogens.

Download/ crawl/ collect data from all the sources & convert data from original sources (Webpages, pdf files, CSV files, …) to structured data fields:

Explain the process chosen and what are all the challenges taken?

Webpages (*Wikipedia Extract):*

1. Using Requests package, webpage is loaded and contents is fetched.
2. Using BeautifulSoup package, all references of the 'table' tag is found on the webpage using find_all function and the table of interest is accessed using the 'class' attribute (wikitable in this case).
3. Table headers (th) from HTML is accessed from “strong” tag and in order to extract the column names, we use text method and strip function.
4. Whole table body(tbody) and table row(tr) is looped to find data within a table cell.
5. Data list is converted to data frame and is converted to csv file.

KEGG database:

1. Since, the diseases in webpage is categorized according to organs in a tree format, the URLs had to be crawled before extracting information of attributes.

   Challenge: We tried extracting the URLs by defining different classes like download URL, get linked URLs, adding URL to visit the webpage and crawl using libraries like logging, time, BeautifulSoup and url_normalize(function of urllib3) in Python. While running the crawler, it extracted all the links of webpage and was picking lot of redundant data.

   How did we overcome: Using regex and BeautifulSoup packages, we used find_all function in the complete HTML tag of index “/entry/[H0-9]+” to extract the URLs that are in the database and appended those URLs .Then, by using JSON library, all URLs were written in a file, post which, the JSON file is closed.
2. Using the above file containing all URLs, all column names with their content in webpage had to be placed in structured format.

   Challenge: For some of the URLS, there was data pertaining to both “Drugs” and “Carcinogens”, in some URLs, there was information regarding either “Drugs” or “Carcinogens” . So, by just extracting based on table headers in HTML tag and using index, for few links, data for “Carcinogens” appeared in “Drugs” column.

   How did we overcome: Using regex and BeautifulSoup packages,we have extracted all “Drugs” that were present in the “Drugs”, using find_next_siblings function,text method and strip function, we extracted all content pertaining to that column.

   Similar process is done for “Carcinogens” and “Gene” as well.
3. **Overall Process:** Above JSON file containing all URL’s of each disease listed in the webpage was called and column names were defined as keys with values as empty list of a dictionary. A loop is run where all table header(th) tags is called and using find_next_siblings, the text is stripped for each of the column.
4. Data list is converted to data frame and is converted to csv file.

Mayo Clinic:

1. In this website, the diseases are organized alphabetically and there is lot of text available for each of the diseases. First, the URLs were crawled before extracting information of attributes.

   While extracting the URLs, we extracted it alphabetically through regex and had to append it to the list and running the crawler writing it back to a JSON file.

   Challenge: We faced similar challenge as Kegg database of table contents getting overlapped.

   How did we overcome: Using regex and BeautifulSoup packages, table header where string is equal to column header is identified and then, using find_next_siblings function, text is stripped for that column. Similar process is done for all attributes as well.
2. **Overall Process:** Above JSON file containing all URL's of each disease listed in the webpage was called and column names were defined as keys with values as empty list of a dictionary. A loop is run where all table header(th) tags is called and using find_next_siblings function, the text is stripped for each of the column.
3. Data list is converted to data frame and is converted to csv file.

Data cleaning/pre-processing as needed:

What data cleaning/ pre-processing techniques were taken up for this stage and why did you feel the requirement for it?

1. Two sets of data are extracted– Structured and Unstructured data.
2. Using Python, all structured datasets from different sources were merged.
   a. Duplicate row entries of same diseases from different sources of Wikipedia were analysed and if it had additional information in column, those were added as additional columns for that disease. Post which, duplicate row entries were dropped.
   b. Duplicate column entries of same diseases from different sources of Wikipedia were analysed and if it had additional information in rows, those were added as additional information in rows for that disease. Post which, duplicate column entries were dropped.
3. For Kaggle dataset, there were multiple csv files so those datasets had to be merged together considering primary key as disease and symptoms.
4. Unstructured dataset of Mayoclinic and Kegg is merged and finally, it is merged with structured dataset and duplicates is removed. Quite a few columns had noise data, which were cleaned to maintain sanity of dataset.
5. Missing data was handled by filling it with blank spaces.

Observations/ Insights and Analysis on the data collected

Exploratory Data Analysis was done on pre-processed dataset. As far as our final dataset is concerned, we observed that most of the attributes were categorical in nature, due to which, we did not see any scope to perform exploratory data analysis to get statistical results and get insights out of it.

Strategy to enhance the data with crowd sourcing methods:

What strategy did you implement to improve on the crowd sourcing methods and why?

1. During pre-processing, we realized that there are multiple entries of disease with variations. Since, we do not have domain expertise, we would need recommendation from medical experts to validate these datasets. This can be done by gathering credible information via surveys in hospitals and medical colleges through medical fraternity.
2. In the final dataset, we have attributes of genes which can be used in further genetic studies.
3. Data on additional diseases or missing attributes can be sourced through surveys in universities or hospitals.
4. Additional data on missing attributes can be sourced from disease registries focused on specific types of diseases (for example the CORONA registry in the US has data on various inflammatory diseases like rheumatoid arthritis and IBD).
5. Numerous university databases exist for specific diseases or groups of diseases, that can be used to enrich this database.

References and Sources used for this Assignment

Please find the links below which were used as references and sources for this assignment:

1. **Structured Sources:**

   Wikipedia:

   https://en.m.wikipedia.org/wiki/List_of_infectious_diseases

   https://en.m.wikipedia.org/wiki/List_of_human_disease_case_fatality_rates

   https://en.m.wikipedia.org/wiki/List_of_genetic_disorders

   Other Websites:

   Table of Diseases | PHA Infection Control (niinfectioncontrolmanual.net)

   Table: Infectious Diseases | HealthEd

   Kaggle:

   https://www.kaggle.com/itachi9604/disease-symptom-description-dataset

2. **Unstructured Sources:**

   Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/index?letter=A

   KEGG database: https://www.genome.jp/brite/br08402