Capstone Project – Fall 2023

PH 1975 – Introduction to Data Science

Dr. Cao

Group Members:

Aaron Niecestro

Jack Mittenthal

Yukuan Pan

Jung Yang

Table of Contents

Section 1) Program Design

We developed a collaborative Python program within Jupyter Notebook. Our process initially began by specifying HIV-related articles in PubMed. We extracted the titles, abstracts, publication times, and author names for all articles containing the keyword “HIV” in the abstract. We utilized the abstracts rather than the titles largely because the abstracts are more likely to contain relevant information. It should be noted that although an article may revolve around “HIV”, there is no guarantee that the title would reflect this.

The following Python packages were used: BioPython, time, csv, pandas, requests, BeautifulSoup, re, tqdm, numpy, matplotlib, plotly, tabulate, and sqlite3.

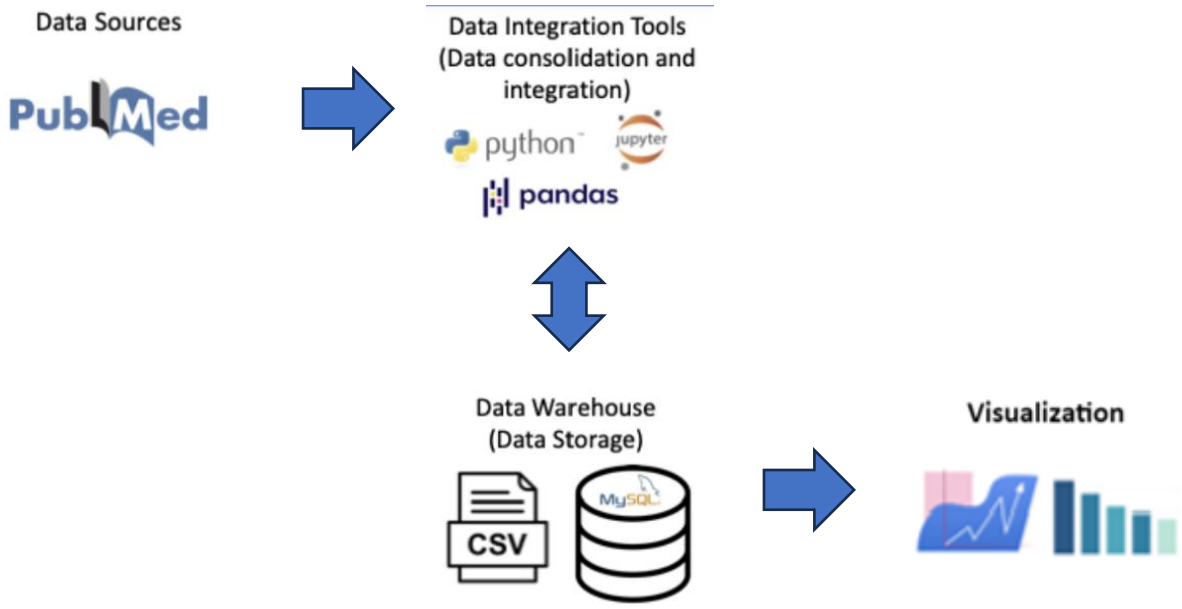
We wrote a function to compile the appropriate article titles, abstracts, publications, and author names into a singular data frame. Upon completion, this file was later converted into a CSV file. The CSV file was saved into the Python environment. Before continuing, we cleaned our dataset to reflect the project’s pre-specified time window of 01/01/2020 – 08/30/2020. Upon completion of data cleaning, we saved a new CSV file to reflect the improvements.

We converted our cleaned file into a new data frame in Python and then proceeded to create an SQLite database. The data frame was loaded into the SQLite database for further analysis. Within the Python environment, we utilized SQL to query publications by individual author names and output only the article titles. For the entirety of the SQL process, a connection was established and later closed once we had finished our queries.

To produce our visualizations, our data frame was loaded back into Python. To display the total number of publications per month, we created a bar chart that shows the monthly outputs as well as an average benchmark for comparison. Annotations for the lowest, highest, below, and above-average months are provided in the plot. Furthermore, the following summary statistics are also available in this report in a tabular format: total, mean, minimum, maximum, standard deviation, first quartile, median, and third quartile. Additional visualizations to see trends over time were provided using a line plot. A benchmark for the average across the entire time window is provided in the same line plot.

Table 1: Workflow Diagram

Table 2: Tool Architecture



Section 2) Implementation Details

Part 1: Scraper Module

The scraper module was specifically designed to extract all publications related to HIV that were published between 01-01-2020 and 08-30-2020. All relevant articles were initially stored in a list, which was then converted to a data frame, which was finally developed into a data frame

1. Specify the keyword “HIV”
2. Set the start and end dates as “2020/01/01” and “2020/08/30” respectively
3. Load the “Biopython” package.
4. Create an empty list to hold retrieved data from PubMed and name it “data”
5. Create and run the customized function “Q1” that does the following:
    a. Display personal email to NCBI before accessing PubMed
    b. Enter PubMed and scrape all available publications and articles related to HIV that were published between the pre-specified start and end dates.
    c. Fetch the following attributes from the appropriate articles and publications: “Title”, “Author List”, “Publication Time”, and “Abstract”
    d. Append the “Title”, “Author List”, “Publication Time”, and “Abstract” information from the articles/ publications into the “data” list
    e. Convert our list “data” into the data frame “df”
    f. Convert the data frame “df” into a CSV file and name the file “pubmed_capstone.csv”

**The scraper module is very intensive and may require a significant amount of time to run. Depending on your device, it is recommended to charge your device, disable sleep mode, and allow the scraper module to run in the background. Run times can take as long as 2 hours.

Part 2: Data Cleaning Module

The CSV file “pubmed_capstone.csv” was thoroughly examined and cleaned for any possible errors, missing values, as well as observations not pertinent to our project. The following steps are provided below the detailed process of cleaning our dataset before the analysis.

1. Read the CSV file “pubmed_capstone.csv” into the Python environment as a data frame
2. Check the dimensions of the data frame
3. From the data frame, extract the following variables “Year”, “Month”, “Month Name”, and “Day”
4. Change the variables listed in step 3 to string variable columns
5. Convert the “pub_time” column to a date/time column
6. Apply the time filter window between “2020-01-01” and “2020-08-30”
7. Sort the dataset to check for any potential mistakes
8. Check the dimensions of the data set again for any missing observations
9. Convert the finalized data frame into a CSV file named “pubmed_capstone2.csv”

Part 3: Database Module

The file from Python must be loaded into an SQLite environment for further testing. The commands provided in the steps below will allow the user to manually enter the names of any publication/ article authors to produce a list of relevant work listed in PubMed.

1. Load the “pubmed_capstone2.csv” file into the environment as a pandas data frame
2. Connect to SQLite and build a database named “pubmed.db”
3. Convert the data frame into a table in SQLite
4. Read the first query to test if the connection to SQLite
5. If successful, load the next commands and run
    a. If not successful, repeat previous steps and troubleshoot for code errors
6. Load the command that allows the user to search all publications based on the author's name.
    a. If successful, the publications will be displayed
    b. If not successful, the publications will not be displayed
7. Close the SQLite connection

Part 4: Visualization Module

We provided charts and figures as supplementary material. These outputs will help the user to better understand the results generated from our code.

1. Load the cleaned CSV dataset as a pandas data frame into the environment
2. Run the following cells to produce specific visualizations and figures:
    a. Cell 2: Bar chart
    b. Cells 3 – 6: Summary statistics
    c. Cell 7: Summary statistics table
    d. Cell 8: Organized dataset based on calendar months.
    e. Cell 9: Line graph

Section 3) Results

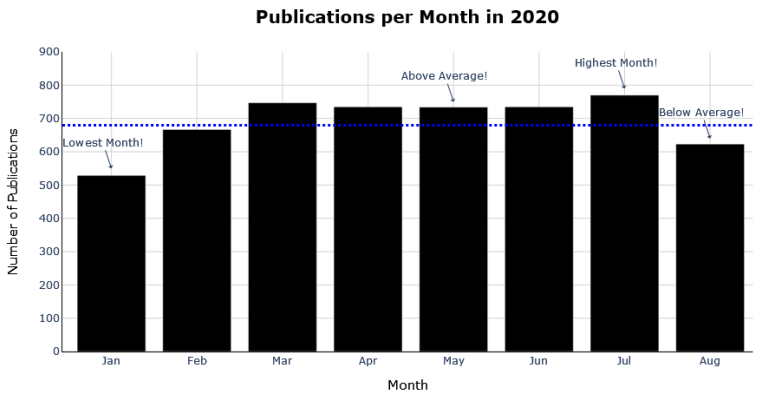
Figure 1. Manual search for stored publications for a specific author

```
[('High sleep-related breathing disorders among HIV-infected patients with sleep complaints.',),
 ('HCV co-infection among people living with HIV Is associated with Higher fracture risk.',)]
```

We produced our SQL database from our finalized list of publications in Python. We can search for any author's name and produce a list of publications relevant to that specific author published between 01-01-2020 and 08-30-2020. See the above sample we provided as an example of our code's output. Figure 1, provided above, is a sample of our code output and shows all HIV-related articles published by author Cheng-Chun Chen between 1/1/2020 and 8/30/2020.

Based on our cleaned dataset, we produced a bar chart reflecting the total number of HIV-related publications available in PubMed over 8 months. The dotted blue line in both the bar and line charts are reflective of the overall average number of publications from January to August in the year 2020.

Figure 1. Bar chart of monthly publications



The greatest number of publications was recorded in July (770 publications), whereas the lowest month was January, with only 529 publications. All but three months in the 8-month time window saw above-average numbers of monthly publications. January, February, and August were the three months that saw below-average numbers of monthly publications.

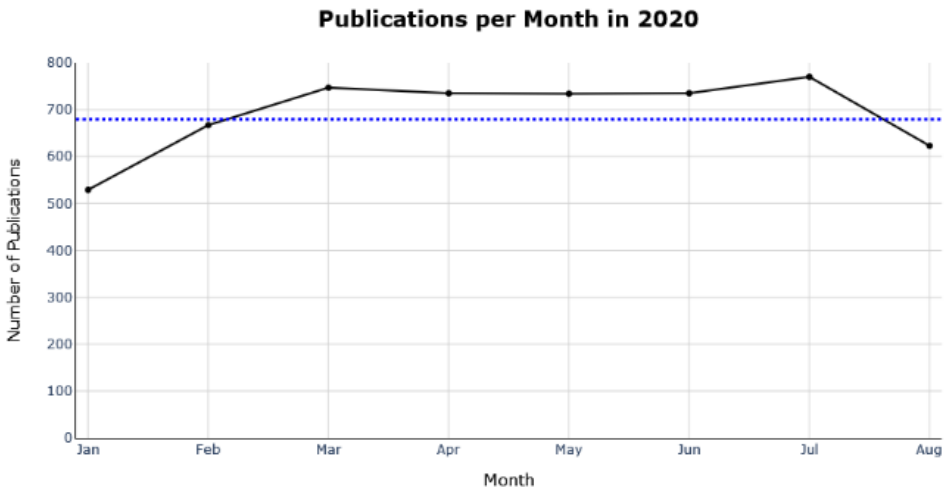
Figure 2. Summary statistics table

Total	Mean	Min	Max	Standard Deviation	IQR 25%	Median	IQR 75%
5540	692.5	529	770	81.5248	656	734.5	738

Overall, the total number of publications over the 8 months was 5,540. The average number of publications was approximately 693 publications, with a standard deviation of approximately 82

publications. The median number of publications was recorded at approximately 735, while the 1st and 3rd quartile were 656 and 738 publications respectively.

Figure 3. Trend line



Based on the trends, we can see that the overall number of publications tends to increase starting from the beginning of the year in January and slowly tapers off beginning in March. From March, the overall number of publications remains relatively steady and above average until July, when the number of publications begins to decline up to August.

Section 4) User Manual/ Guide

1. Part 1: Scraper Module

    a. Install Anaconda
    b. Open Anaconda
    c. Open the Anaconda prompt and install the “BioPython” package.
    d. Load “BioPython” package
    e. Upon completion, restart Anaconda
    f. Open Jupyter Notebooks
    g. Download “PubMed_Scraper.ipynb”
    h. Upload the “PubMed_Scraper.ipynb” file into the Documents folder in Jupyter Notebooks
    i. Open the “PubMed_Scraper.ipynb” file in Jupyter Notebooks environment.
    j. Run and load the packages cell [Jupyter Cell 1]
    k. Run the scraper function (function Q1) Jupyter cell 2
        i. This function requires approximately 2 hours to finish. Therefore, please let the function run in the background and disable your device’s sleep function. Do not minimize while the application is running. Please keep your computer charging during this process.
    l. Run Jupyter cell 3 to call the Q1 function

2. Part 2: Data Cleaning

    a. Load the CSV dataset as a pandas data frame
    b. Check the dimensions of our data frame
    c. Extract the following variables from our “pub_time” column: “year”, “month”, “month_name”, “day”
    d. Change the following columns “title”, “month_name”, “author_list”, and “abstract” to string columns
    e. Convert the “pub_time” column to a date/time column.
    f. Filter time window between “2020-01-01” and “2020-08-30”
    g. Sort “pub_time” observations to ensure no mistakes were made up to this point
    h. Check the dimensions of our data frame to check for missing data and observations
    i. Convert the data frame to another CSV file and name it “pubmed_capstone2.csv”

3. Part 3: Building SQL Database

    a. Load the cleaned CSV dataset as a pandas data frame into the environment.
    b. Establish the initial connection to SQLite and automatically build a database named “pubmed.db”
    c. Convert the data frame to a table in SQLite
    d. Read the first SQL query to ensure proper function
        i. If successful, the first five observations will be displayed
        ii. If not successful, review previous steps
    e. Load the command that would allow one to enter a specific author’s name
    f. Load the next SQL query
        i. If successful, the author’s titles will be displayed

        ii. If not successful, review previous steps
    g. Close the SQLite connection

4. Part 4: Visualization

    a. Load the cleaned CSV dataset as a pandas data frame into the environment.
        i. Check to make sure the upload was successful
    b. Run the second cell of the Visualization module to create the bar chart.
    c. Run cells 3 – 6 of the Visualization module to create the summary statistics
    d. Run cell 7 of the Visualization module to create a table of the summary statistics
    e. Run cell 8 to organize the dataset to follow the calendar months
    f. Run cell 9 to produce the trend/ line graph

**If there were any errors, please refer to the code as well as the reference material for further guidance