Capstone Project Group 9

Name

Table of Contents

SECTION 1) Program Design ***2***

Summary of tool workflow **2**

Table 1: Tool Workflow **2**

Table 2: Tool Architecture **2**

SECTION 2) Implementation Details ***3***

Crawler Module **3**

Database Module **3**

Visualization Module **4**

SECTION 3) Results ***4***

SECTION 4) User manual/guide ***5***

SECTION 1) Program Design

Summary of Tool Workflow

We constructed a Python program in a collaborative Jupyter Notebook file. We started by building a crawler module using the BioPython and Pandas python packages, writing a function to collect information on papers from PUBMED and save the output to a CSV file. Next,

Table 1: Tool Workflow

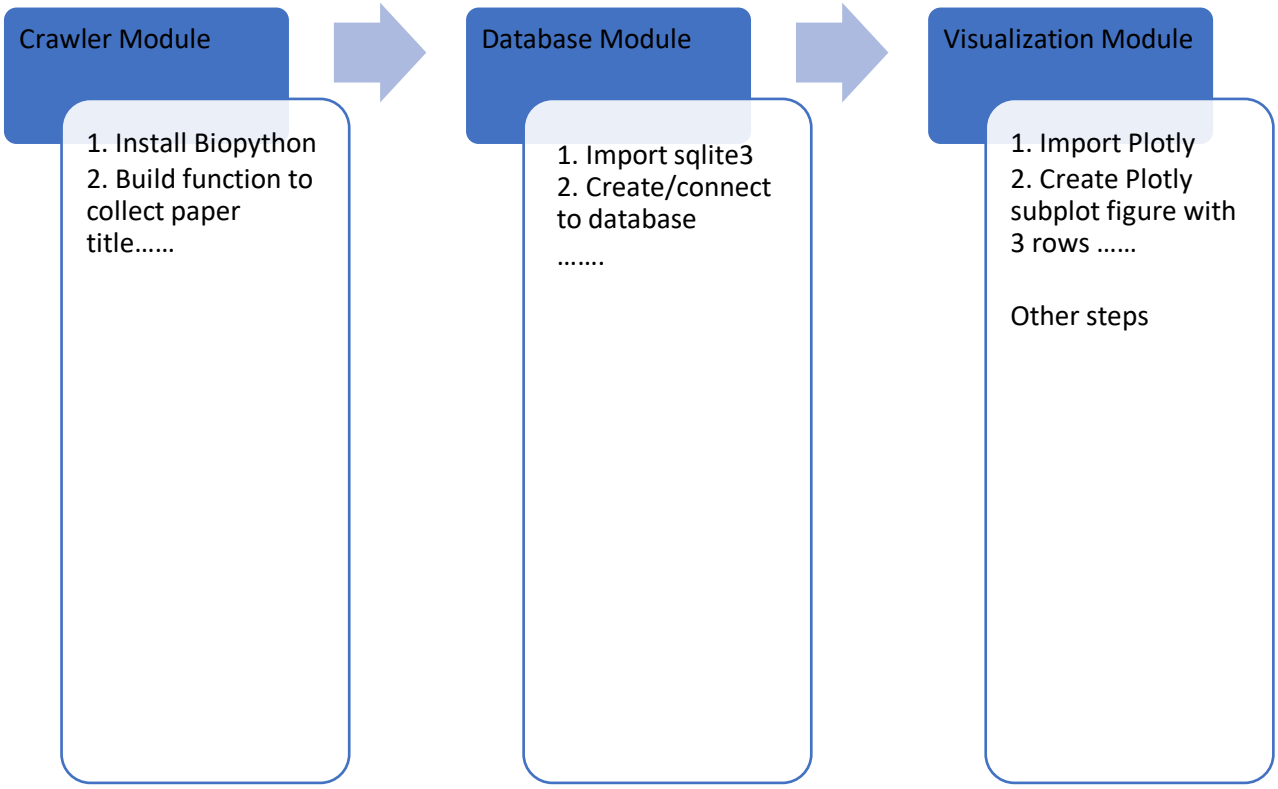
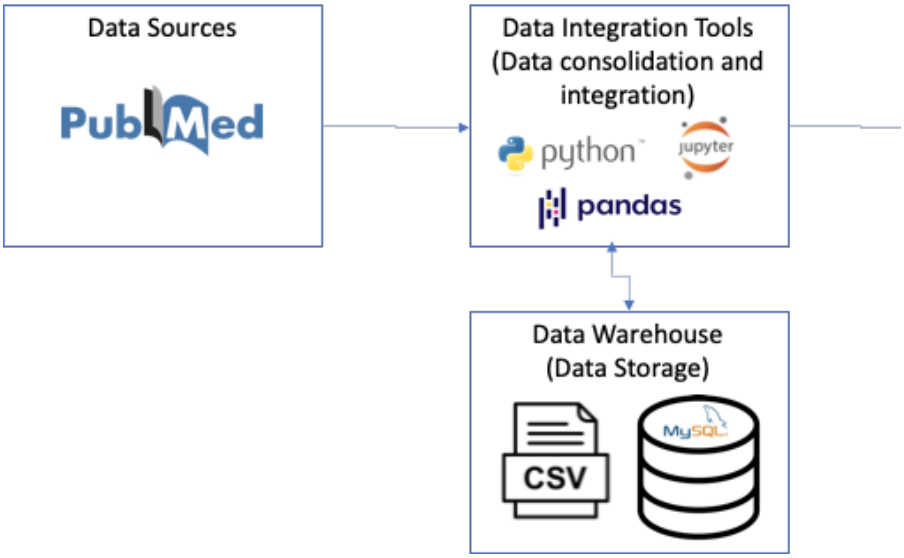


Table 2: Tool Architecture



SECTION 2) Implementation Details

Crawler Module

We follow the process described below using biopython and pandas to create a crawler module that collects paper title, author list, publication time, and abstract from PUBMED for a given keyword (i.e., HIV) within a pre-specified time window (i.e., 01/01/2020 – 08/30/2020), and save the retrieved data to a CSV file.

1. Install Biopython using 'pip install'
2. Other steps
3.
4.
5.

Database Module

We follow the process below using sqlite3 and pandas to create a database module that reads the CSV file created in the Crawler Module to SQLite to build a database automatically. Then we implement SQL code to query the publications by author's name.

1. Import sqlite3
2. Create SQLite database named 'pubmed_crawl.db', or connect to this database if it already exists in the current directory
3. Other steps
4.
5.

Visualization Module

We follow the steps described below using plotly to implement a visualization module that i) reads the CSV file, ii) shows the number of publications in each month, iii) visualizes the trend of the publication numbers over time, and iv) visualizes the summary statistics for the publication number per month.

1. Import plotly.express, plotly.graph_objects, and plotly.subplots in order to create a dashboard to visualize the above data
2. Other steps
3.
4.
5.
6.
7.

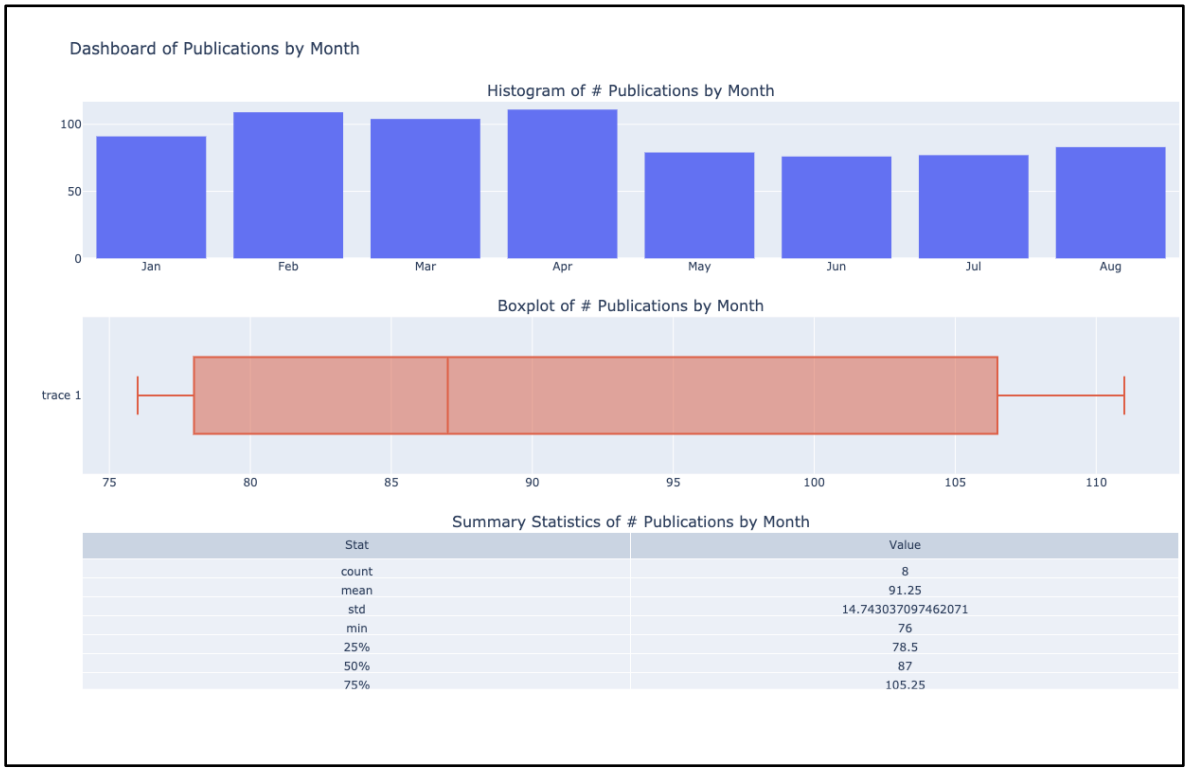
SECTION 3) Results

For the second module in implementing SQL code to query the publications by author's name we did a demo run using the author input name of "Merlin L Robb" and out module returns the number of the articles and article titles for that author.

```
Merlin L Robb has (9,) articles.

('Late boosting of the RV144 regimen with AIDSVAX B/E and ALVAC-HIV in HIV-uninfected Thai volunteers: a double-blind, randomised controlled trial.',)
('HIV vaccine delayed boosting increases Env variable region 2-specific antibody effector functions.',)
('A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection.',)
('Molecular dating and viral load growth rates suggested that the eclipse phase lasted about a week in HIV-1 infected adults in East Africa and Thailand.',)
('Safety and immunogenicity of Ad26 and MVA vaccines in acutely treated HIV and effect on viral rebound after antiretroviral therapy interruption.',)
('Dynamic MAIT cell response with progressively enhanced innateness during acute HIV-1 infection.',)
('HIV status disclosure by Nigerian men who have sex with men and transgender women living with HIV: a cross-sectional analysis at enrollment into an observa
('Continuous prophylactic ARV/ART since birth reduces seeding and persistence of the viral reservoir in vertically HIV-infected children.',)
('Preferential infection of α4β7+ memory CD4+ T cells during early acute HIV-1 infection.',)
```

Our first module extracted a total of 730 pubmed articles for the query search of "HIV" with publication date between 01/01/20 and 08/30/20. Below is the dashboard we created using plotly to visualize our results from module 1.



Summary Statistics of # Publications by Month

Stat	Value
count	8
mean	91.25
std	14.743037097462071
min	76
25%	78.5
50%	87
75%	105.25

SECTION 4) User manual/guide

1. Download the “Group 9 Project Code.ipynb” file and save into the folder where you will be working from.
2. Open Anaconda.
3. Launch Jupyter Notebook.
4. Other steps
5.
6.
7.
8.