Capstone Project Group 13

Chunhui Gu, Ruoxing Li, Xun Xu, Feng Zhang

**Table of Contents**

[***SECTION 1) Program Design 2***](#_heading=h.3znysh7)

[**Summary of tool workflow 2**](#_heading=h.2s8eyo1)

[**Table 1: Tool Workflow 2**](#_heading=h.17dp8vu)

[**Table 2: Tool Architecture 2**](#_heading=h.2xhcjkjqkjmt)

[***SECTION 2) Implementation Details 3***](#_heading=h.3rdcrjn)

[**Crawler Module 3**](#_heading=h.26in1rg)

[**Database Module 3**](#_heading=h.lnxbz9)

[**Visualization Module 4**](#_heading=h.35nkun2)

[***SECTION 3) Results 4***](#_heading=h.1ksv4uv)

[***SECTION 4) User manual/guide 5***](#_heading=h.hk2qt552aniw)

# 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

Crawler Module

1. Install Biopython

2. Build function to collect paper title……

Database Module

1. Import sqlite3

2. Create/connect to database

…….

Visualization Module

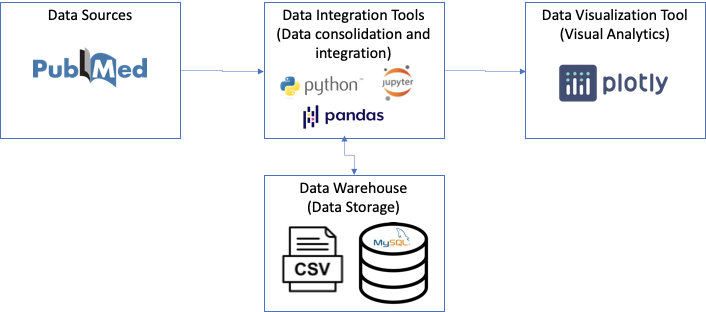
1. Import Plotly

2. Create Plotly subplot figure with 3 rows ……

Other steps

## 

## 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

## 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 ‘HIV.db’, or connect to this database if it already exists in the current directory
3. Other steps

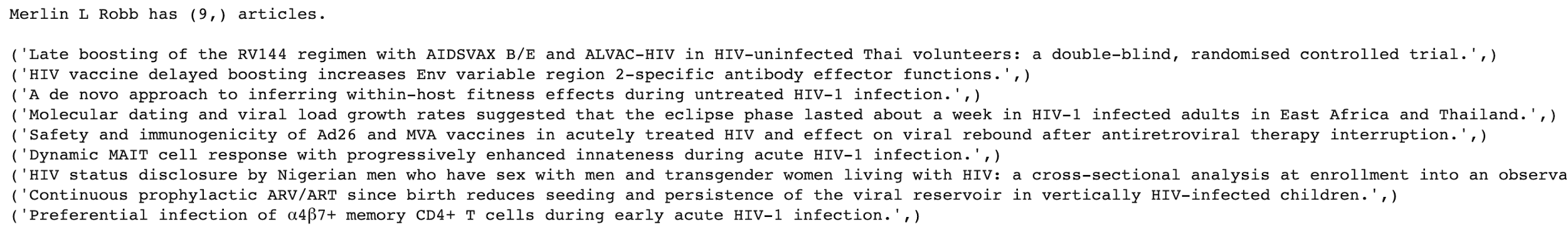
## Visualization Module

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.

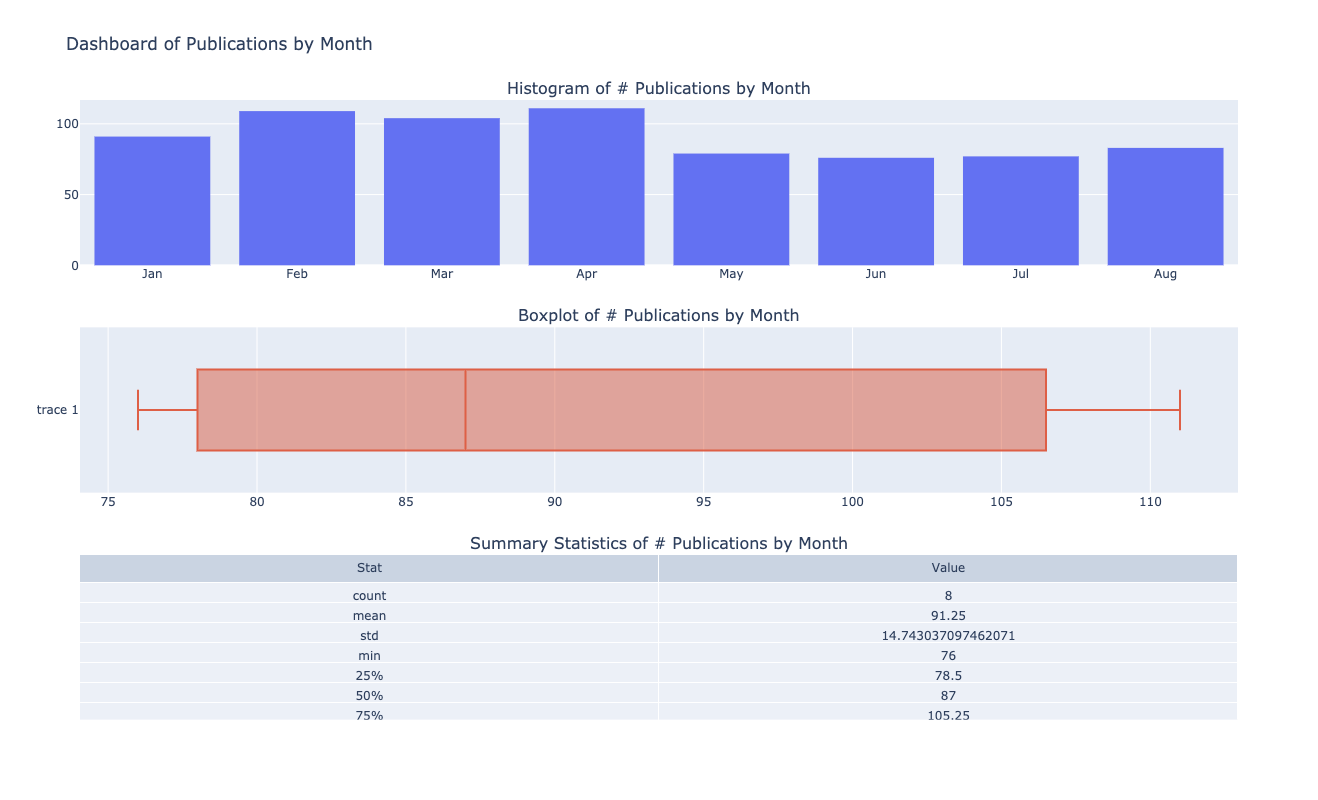
1. Import plotly.express, plotly.graph\_objects, and plotly.subplots in order to create a dashboard to visualize the above data
2. Other steps

# 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.



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.



# 

# 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