|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  |
| Capstone Project Report  PH1975\_Intro to DS  Jiahui Jiang, Miao Tang, Weilu Zhao, Jiyuan Tang | | |
|  | | December 9th, 2022 |

Table of Contents

[Section 1) Introduction of oral squamous cell carcinoma 2](#_Toc121438171)

[Section 2) Program Design 2](#_Toc121438172)

[Table 1: Workflow Diagram 3](#_Toc121438173)

[Table 2: Tool Architecture 3](#_Toc121438174)

[Section 3) Analysis Details 4](#_Toc121438175)

[Crawler Module 4](#_Toc121438176)

[Database Module 4](#_Toc121438177)

[Visualization Module 4](#_Toc121438178)

[Section 4) Analysis Results 5](#_Toc121438179)

[Section 5) User Manual 7](#_Toc121438180)

# Section 1) Introduction of oral squamous cell carcinoma

Almost all the cancers in the oral cavity and oropharynx are squamous cell carcinomas, also called squamous cell cancers. These cancers start in squamous cells, which are flat, thin cells that form the lining of the mouth and throat. Over 95% of people with oral squamous cell carcinoma smoke tobacco, drink alcohol, or both. Because the more and more use of tobacco and alcohol in our daily lives, the prevalence of oral squamous cell carcinoma turned to higher and higher.

In order to prevent the oral squamous cell carcinoma from more serious situation, it requires early detection by screening. Then, by proper utilization of treatment such as surgery, radiation, or both, the 5- year survival rate became higher recently.

Because the current situation of oral squamous cell carcinoma and the prevalence of oral squamous cell carcinoma, more and more researchers and scholars are trying hard to study oral squamous cell carcinoma in any perspectives and deep. They want to find a way to help patients and human beings completely cured from oral squamous cell carcinoma.

Due to more and more scholars did research and studies in oral squamous cell carcinoma, there are many valuable articles have been published previously each month. We are conducting the research on Pubmed to figure out how many articles have been published alone with authors and some other associated details for period between 1/1/2020 and 8/30/2020 with keyword ‘Squamous Cell Carcinoma’.

# Section 2) Program Design

We constructed a Python program in a collaborative Jupyter Notebook file. We started by building a crawler module using the BioPython, Pandas, bs4, and pubmed\_parser. Beautiful Soup is for pulling data out of HTML and XML files, and pubmed\_parser is for parsing the PubMed open-access subset information into our Python dictionary for research. We are writing a function to collect information of published articles from PubMed between 1/1/2020 and 8/30/2020 regarding the keyword ‘Squamous Cell Carcinoma’ and save the outputs to a CSV file. We included those articles’ PMID, title, author list, publication year, publication month, and article abstract.

Next, we used pathlib to help classes representing filesystem paths with semantics appropriate for different operating systems. We imported sqlite3 and then import our generated CSV file containing articles’ information as the database. Then by manually input author name, it would implement SQL code to find out associated his/her publications.

Then, we imported the Plotly module, the plotly.express can help create entire figures at once, the plotly.io is for help construct figures; Additionally, we utilized the seaborn and matplolib.pyplot, both modules are based on matlab and provide the high-level interface for constructing statistical graphics, charts, tables and etc. We generated the monthly publication table, monthly publication line chart, monthly publication bar chart, the vertical violin plot, and box plot.

## Table 1: Workflow Diagram

A picture containing diagram

Description automatically generated

## Table 2: Tool Architecture

Diagram

Description automatically generated

# Section 3) Analysis Details

## Crawler Module

We follow the process described below using biopython and pandas to create a crawler module that collects PMID, title, author list, publication year, publication month, and article abstract from PubMed for a keyword, we are using ‘Squamous Cell Carcinoma’ as our demo keyword. With pre-specified time window 1/1/2020 – 8/30/2020, then save the retrieved information and data to a cvs file named ‘capstone.csv’.

1. Install biopython using ‘pip install’
2. Import pandas, BeasutifulSoup, and pubmed\_parser
3. Enter the keyword, ‘Squamous Cell Carcinoma’
4. All data saved in ‘capstone.csv’

## 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 ‘articles.db’
3. Input the demo author name, J Claveau
4. Then the data contains such author’s article alone with its details shown

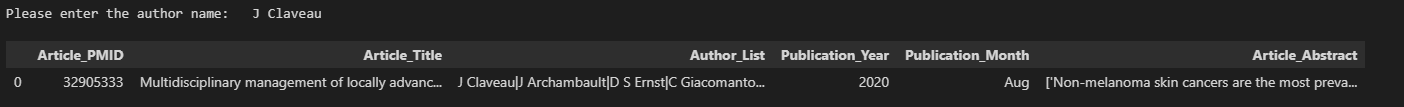
## Visualization Module

We follow the steps described below using plotly, matplotlib.pyplot, and seaborn 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 by month, iv) visualizes the summary statistics for the publication number per month and clearly visualizing the difference between each month, and v) visualizes the details of time frame 1/1/2020 – 8/30/2020 by utilizing box plot and violin plot.

1. Import plotly.express and plotly.io to create a dashboard to visualize the above data
2. Import matplotlib.pyplot and seaborn for utilizing matlab and provide the high-level interface for constructing statistical graphics
3. Import previously generated csv file ‘capstone.csv’
4. Construct monthly publication table
5. Construct monthly publication line chart
6. Build monthly publication bar chart
7. Build the vertical violin plot and box plot.

# Section 4) Analysis Results

We did the demo run for keyword ‘Squamous Cell Carcinoma’ and author’s name J Claveau. And the result shows as the following



Our first module extracted a total of 5368 PubMed articles for the keyword ‘Squamous Cell Carcinoma’ with publication date between 1/1/2020 and 8/30/2020. Below are the dashboard using plotly, matplotlib.pyplot and seaborn to visualize the results.

Chart, line chart

Description automatically generated

A picture containing text, clipart, screenshot

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

By using simple counts function we can get following results:

A picture containing text, screen, scoreboard

Description automatically generated Table

Description automatically generated with medium confidence

# Section 5) User Manual