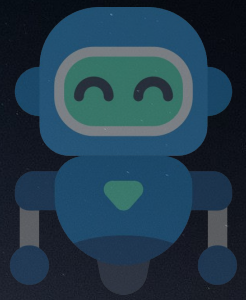
Custom Bot Project

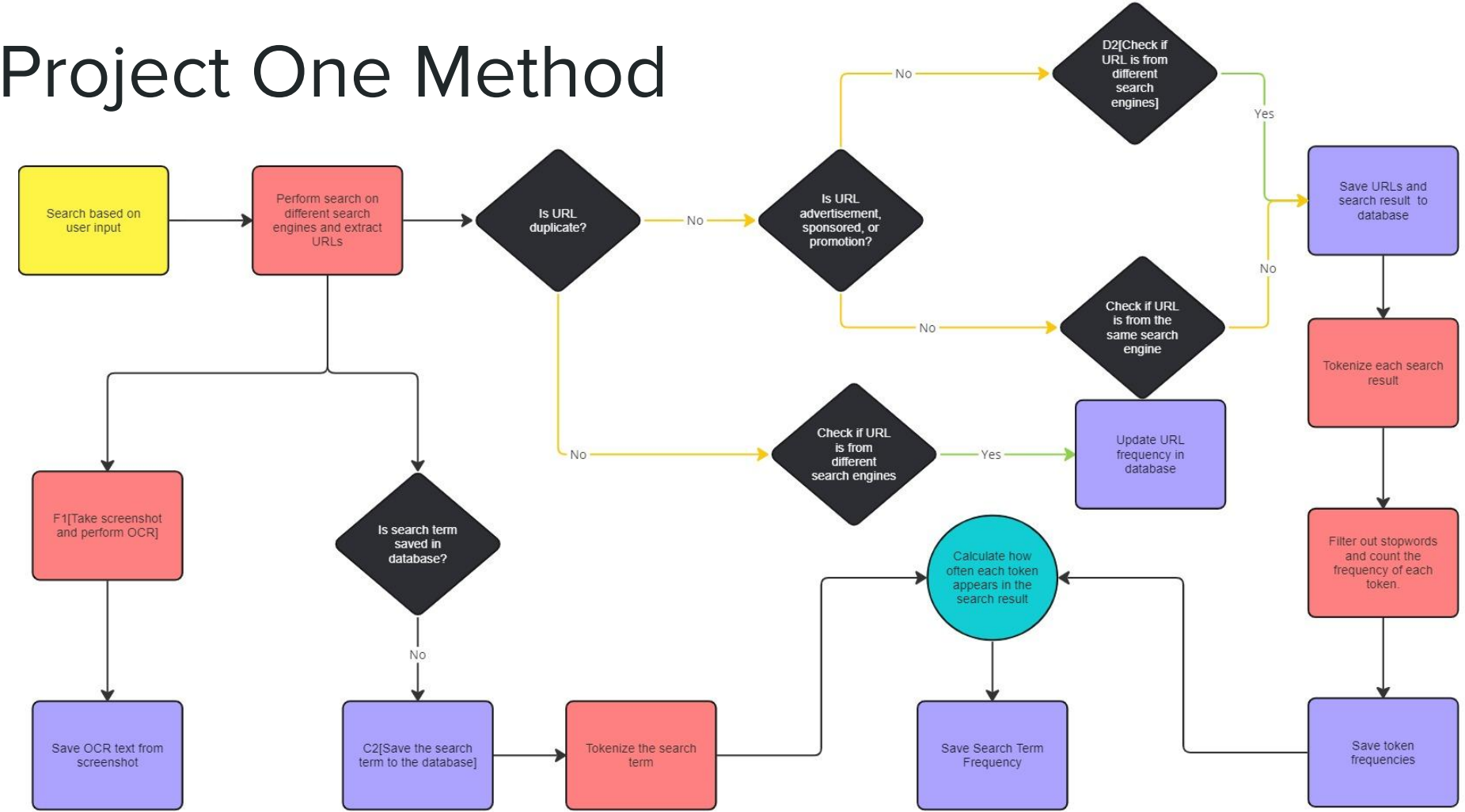


Project 1: Goal

Project 1 - Advanced Web Scraping

1. Create a SQL database (MY_CUSTOM_BOT): Design tables to capture search terms and results data.
2. Python Script to perform searches on search engines like- Google, Bing, Yahoo, DuckDuckGo
3. Capture URL Data using OCR tools
4. Remove Ads, Duplicate URLs from Search Results
5. Save URL List to Database
6. Frequency Data Calculation

Project One Method



The Problem or Challenge

Web Scraping Efficacy Factors -

- Google's scroll down code is more effective due to page structure differences, dynamic content loading, and anti-scraping measures.
- High volume scraping from a single IP is not allowed.
- Capturing screenshots during web scraping is challenging due to dynamic content, scrolling, timing, rendering issues, resource-intensive performance, and browser compatibility issues.
- Slow, resources intensive, not scalable
- Unreliable -- breaks when website changes and works poorly with responsive design techniques
- Difficult to parse data
- Often prohibited by TOS

1. Extraction

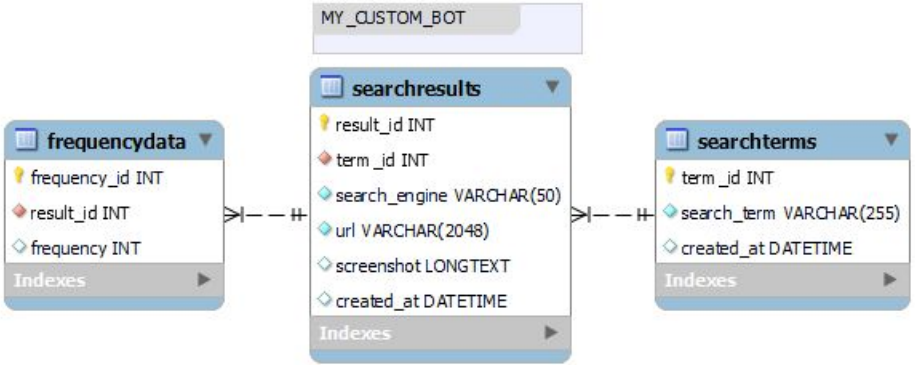
- URL Scraping: Selenium
- URL Screenshot Capture: Selenium
- OCR (Text Extraction from Screenshots): Pytesseract

2. Transformation

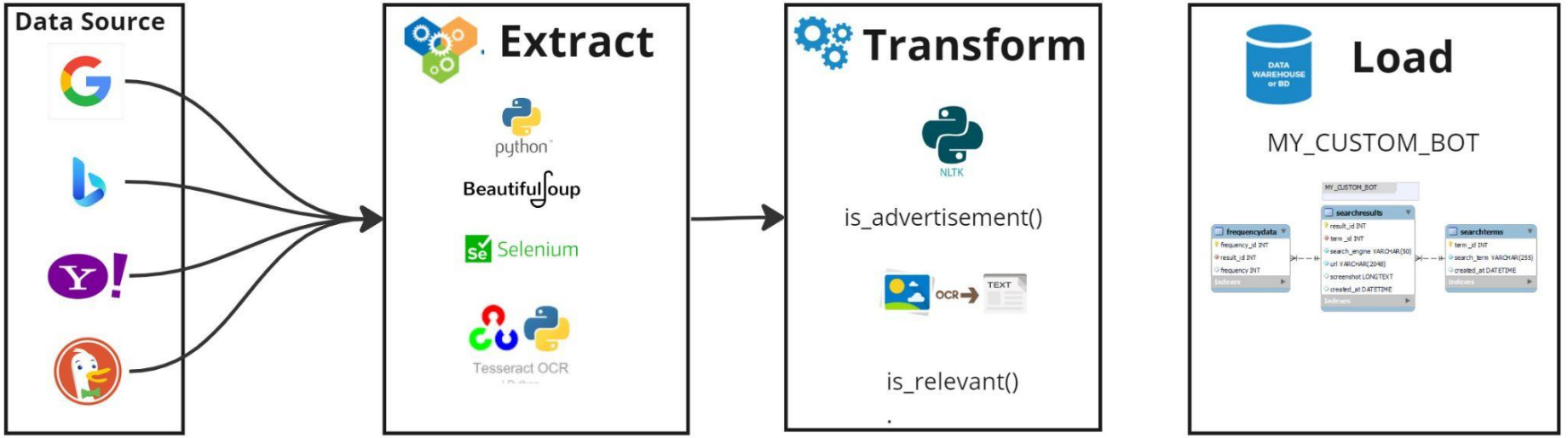
- Search Term
- URLs List
- URLs Number

- NLTK used for tokenize and remove stop word from search term to generate relevant Keyword
- Remove ads, sponsored,promotional URL using Function
- Frequency Count

3. Load



- Search Term - searchterms Table
- URLs- searchresults
- Frequency Count


Data Source
Extract
python
Beautiful Soup
Selenium
Tesseract OCR
Transform
NLTK
is_advertisement()
OCR
TEXT
is_relevant()
.
Load
DATA WAREHOUSE or BD
MY_CUSTOM_BOT
MY_CUSTOM_BOT
frequencydata
frequency_id INT
result_id INT
frequency INT
Indexes
searchresults
result_id INT
term_id INT
search_engine VARCHAR(50)
url VARCHAR(2048)
screenshot LONGTEXT
created_at DATETIME
Indexes
searchterms
term_id INT
search_term VARCHAR(255)
created_at DATETIME
Indexes

Search results

result_id	term_id	search_engine	url	screenshot	created_at
1	1	Bing	https://www.cancer.org/research/acs-research...	C:\Users\mahbu\screenshots\https:__www_ca...	2024-05-09 16:23:36
2	1	Bing	https://www.cancer.gov/news-events/cancer-c...	C:\Users\mahbu\screenshots\https:__www_ca...	2024-05-09 16:23:48
3	1	Bing	https://www.cancerresearch.org/blog/septemb...	C:\Users\mahbu\screenshots\https:__www_ca...	2024-05-09 16:24:00
4	1	Bing	https://medicalxpress.com/news/2023-08-child...	C:\Users\mahbu\screenshots\https:__medicalx...	2024-05-09 16:24:14
5	1	Bing	https://en.wikipedia.org/wiki/Cancer_immunoth...	C:\Users\mahbu\screenshots\https:__en_wikip...	2024-05-09 16:24:27
6	1	Bing	https://www.cancerresearch.org/cancer-types/...	C:\Users\mahbu\screenshots\https:__www_ca...	2024-05-09 16:24:40
7	1	Bing	https://publications.aap.org/pediatrics/article/1...	C:\Users\mahbu\screenshots\https:__publicatio...	2024-05-09 16:24:56
8	1	Bing	https://www.uspto.gov/subscription-center/20...	C:\Users\mahbu\screenshots\https:__www_us...	2024-05-09 16:25:08
9	1	Yahoo	https://www.msn.com/en-us/news/other/immu...	C:\Users\mahbu\screenshots\https:__www_ms...	2024-05-09 16:55:45
10	1	Yahoo	https://www.msn.com/en-us/health/other/supe...	C:\Users\mahbu\screenshots\https:__www_ms...	2024-05-09 16:55:55
11	1	Yahoo	https://www.cancerresearch.org/blog/septemb...	C:\Users\mahbu\screenshots\https:__www_ca...	2024-05-09 16:56:07
12	1	Google	https://siop-online.org/who-global-initiative-for-...	C:\Users\mahbu\screenshots\https:__siop-onlin...	2024-05-09 16:57:47
13	1	Google	https://www.nytimes.com/2023/10/06/opinion/...	C:\Users\mahbu\screenshots\https:__www_ny...	2024-05-09 16:58:05
14	1	Google	https://www.acco.org/us-childhood-cancer-stat...	C:\Users\mahbu\screenshots\https:__www_ac...	2024-05-09 16:58:17

Frequency counts

frequency_id	result_id	frequency
2	2	1
3	3	1
4	4	1
5	5	1
6	6	1
7	7	1
8	8	1
9	9	2
10	10	2
11	11	2
12	12	1
13	13	1
14	14	1
15	15	1

What will I do next?

Next Steps for Web Scraping Project:

Efficacy Factors:
• Explore alternative strategies for scraping search results pages.
• Use techniques to mitigate rate constraints and anti-scraping measures.

Optimizing Screenshot Capture:
• Enhance the screenshot capture process.
• Investigate alternative methods for capturing screenshots.

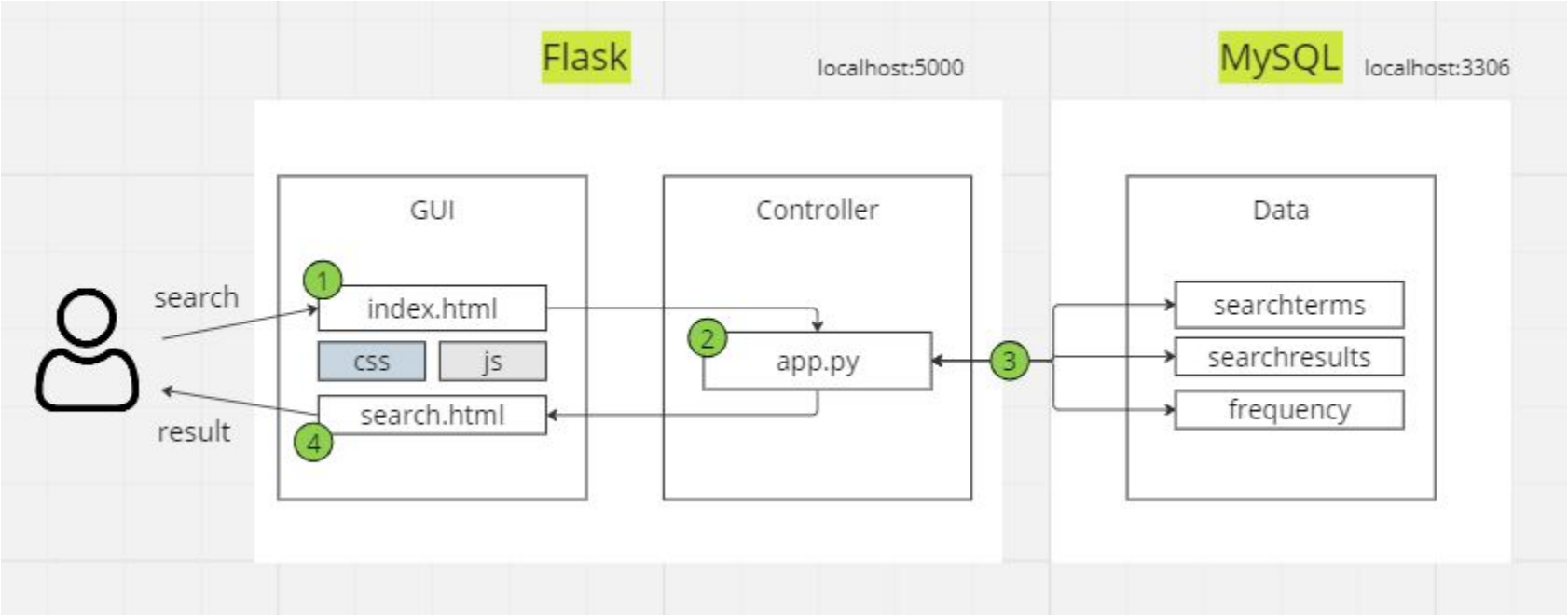
Database Management:
• Refine the database schema and optimize queries.
• Implement data validation and integrity constraints.

Error Handling and Logging:
• Improve error handling mechanisms and logging for troubleshooting and monitoring.

Project 2: Goal

- Build a visually aesthetic web browser GUI that allows users to enter search terms
- Feed the user search terms to your Data Ingestion Engine
- Display 30-100 web URLs ordered by frequency in desc

Method



MVC (Model-View-Controller)

The Problem or Challenge

1. Python on Web

- Remote Host
  - Greenhost
  - Glitch.com
- Local Host
  - plot/dash, javascript
  - Flask web framework

2. Web to MySQL

- mysql-connector
- mysqlclient
- SqlAlchemy

The Output

127.0.0.1:5000/search

Alpha Bot

Search: Submit

Search Results:

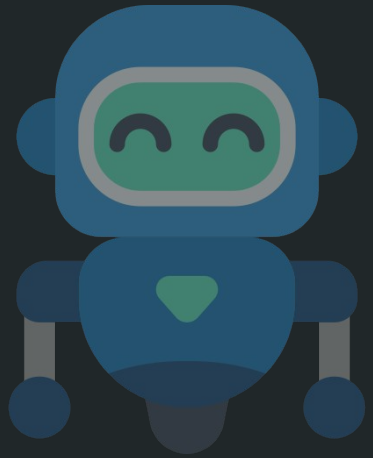
Frequency	URL
63	https://en.wikipedia.org/wiki/Cancer_immunotherapy
63	https://www.acco.org/us-childhood-cancer-statistics/
51	https://www.cancer.gov/types/childhood-cancers/late-effects-hp-pdq
36	https://www.stjude.org/media-resources/news-releases/2021-medicine-science-news/chemoimmunotherapy-dramatically-improved-survival-of-high-risk-neuroblastoma-patients.html
29	https://www.cancerresearch.org/cancer-types/childhood-cancer
27	https://www.childrenscancercause.org/facts
20	https://www.cancer.gov/news-events/cancer-currents-blog/2021/childhood-cancer-accelerating-progress-sharpless
20	https://www.cancerresearch.org/blog/september-2019/immunotherapy-pediatric-cancer-baumeister-q-a
20	https://www.cancer.org/research/acs-research-news/the-current-and-future-promise-of-immunotherapy-for-childhood-cancers.html
18	https://www.stjude.org/inspire/news/new-study-achieves-99-percent-remission-in-children-who-relapse-with-most-common-childhood-cancer.html
18	https://www.who.int/news-room/fact-sheets/detail/cancer-in-children
18	https://www.uspto.gov/subscription-center/2022/uspto-extends-cancer-immunotherapy-pilot-program-until-september-30-2022
17	https://www.hopkinsmedicine.org/inhealth/about-us/immunotherapy-precision-medicine-action-policy-brief
16	https://curesearch.org/understanding-childrens-cancer/childhood-cancer-statistics/5-year-survival-rate/
16	https://www.who.int/news-room/feature-stories/detail/cancer-centres-of-excellence-help-increase-survival-rates-among-children
16	https://www.gao.gov/blog/push-increase-pediatric-research-and-greater-diversity-cancer-treatment-trials
16	https://www.texaschildrens.org/departments/immunotherapy-program

What will I do next?

- How to avoid pop-up windows
- How to scrape result across multiple pages
- How reduce long run time
- How to improve schema structure

Thank you!

To whom gave us instructions!



Group 5: Mahbuba Jyoti, Zhongming Wu

github.com/zwu009/flask-python-search-de