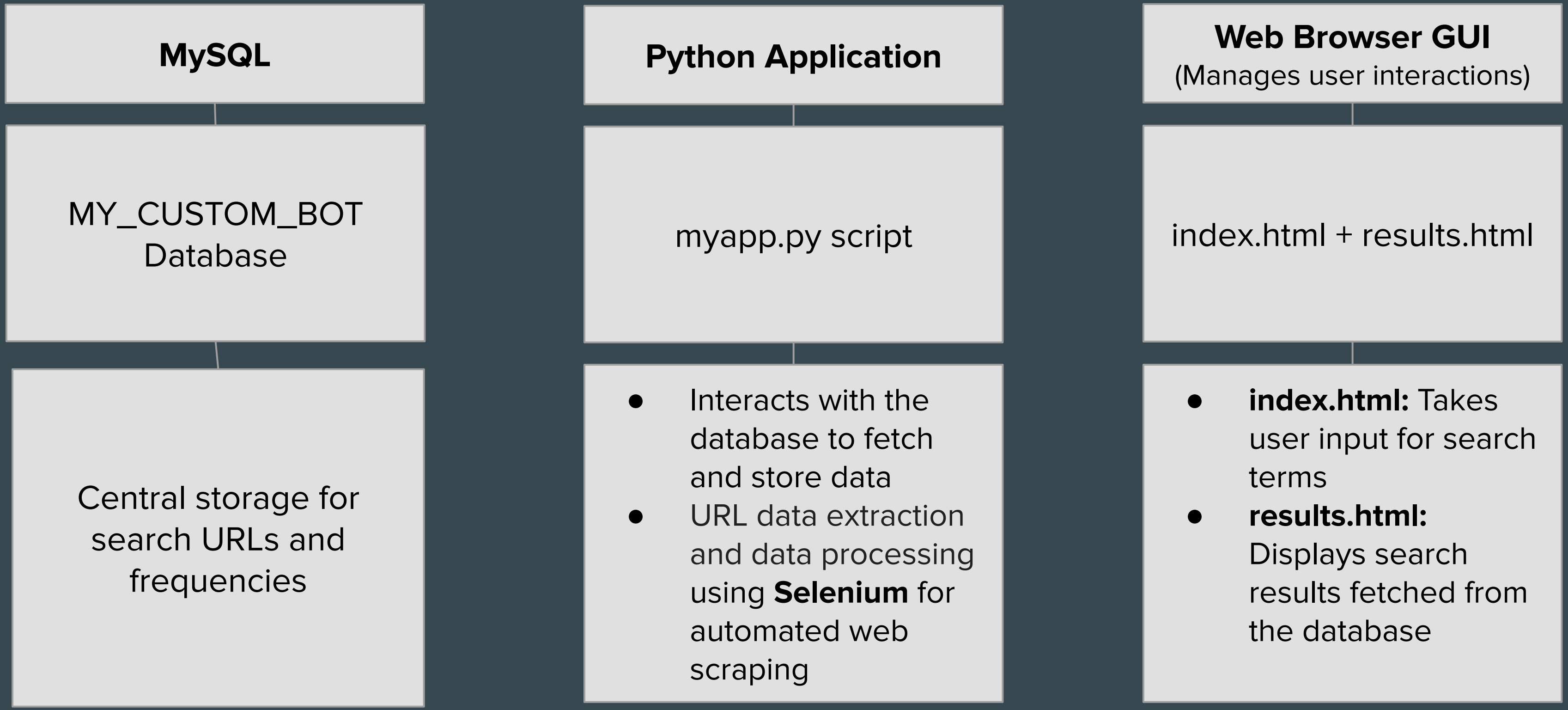
Advanced
Web Scraping + Search

- Artjola Meli
- Shradha Godse

Project Structure



MySQL

MY_CUSTOM_BOT Database

Central storage for search URLs and frequencies

Python Application

myapp.py script

- Interacts with the database to fetch and store data
- URL data extraction and data processing using **Selenium** for automated web scraping

Web Browser GUI
(Manages user interactions)

index.html + results.html

- **index.html:** Takes user input for search terms
- **results.html:** Displays search results fetched from the database

Technologies Used

01	**MySQL - Database**	● Insert the search results ● Store the URLs and Frequency ● Display the URLs and Frequency
02	**Flask Framework**	● Routing and View Functions ● Integration with Selenium & MySQL ● Server Configuration & Error Handling
03	**Selenium**	● Web Browser Automation ● Fetching & Scraping URLs from Search Engines ● Handling Pagination
04	**BeautifulSoup**	● HTML Parsing ● Data Extraction ● Enhancing Data Quality

Duplicate Elimination Logic

Database Primary Key Constraint:

- **Primary Key on URL:**
  - The URL field is set as the primary key in the database.
  - This constraint automatically prevents duplicate URLs from being inserted.

In-Memory Check:

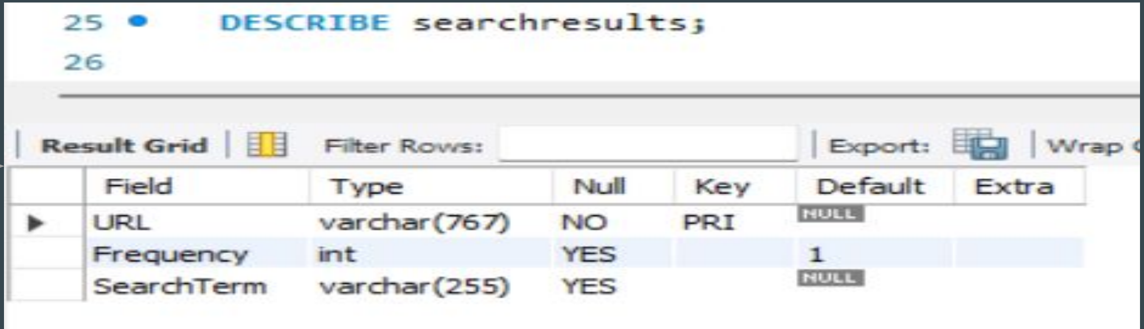
- **Session-Based List:**
  - During each session, maintain a list of URLs already collected and before adding a new URL, check if it already exists.
  - This immediate check ensures no duplicates within the same scraping session.

Advertisement Elimination Logic

- **HTML Tag Selection:**
  - Target tags (<h3>, <h4>) used for organic results.
  - Avoid structures or markers indicating ads.
- **CSS Selector Waiting:**
  - Wait for elements that indicate organic results.
  - Avoid partial results and late-loading ads.
- **URL Pattern Filtering:**
  - Extract URLs with patterns typical for organic results.
  - Skip URLs with ad-related parameters or structures.
- **Ad Marker Exclusion:**
  - Identify and skip elements indicating advertisements.

MY_CUSTOM_BOT Database

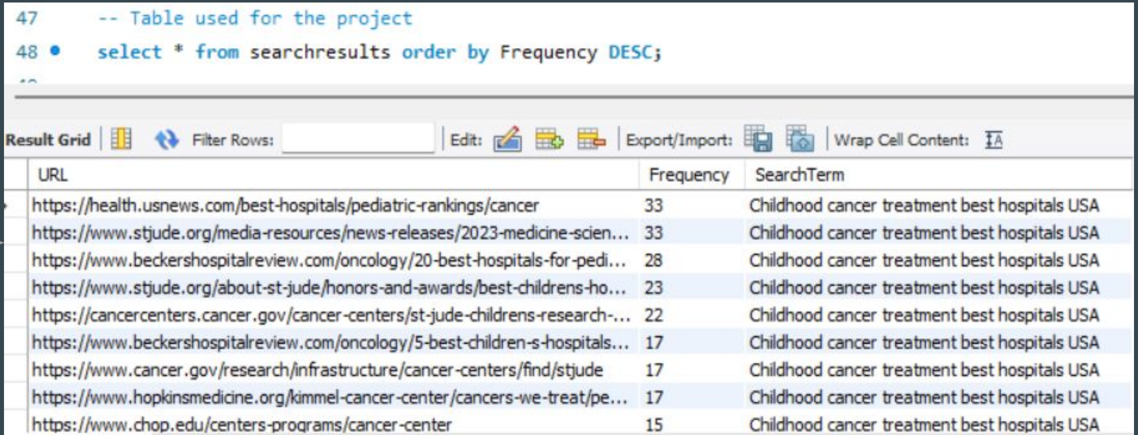
Table Description



```sql
DESCRIBE searchresults;
```

Field	Type	Null	Key	Default	Extra
URL	varchar(767)	NO	PRI	NULL	
Frequency	int	YES		1	
SearchTerm	varchar(255)	YES		NULL	

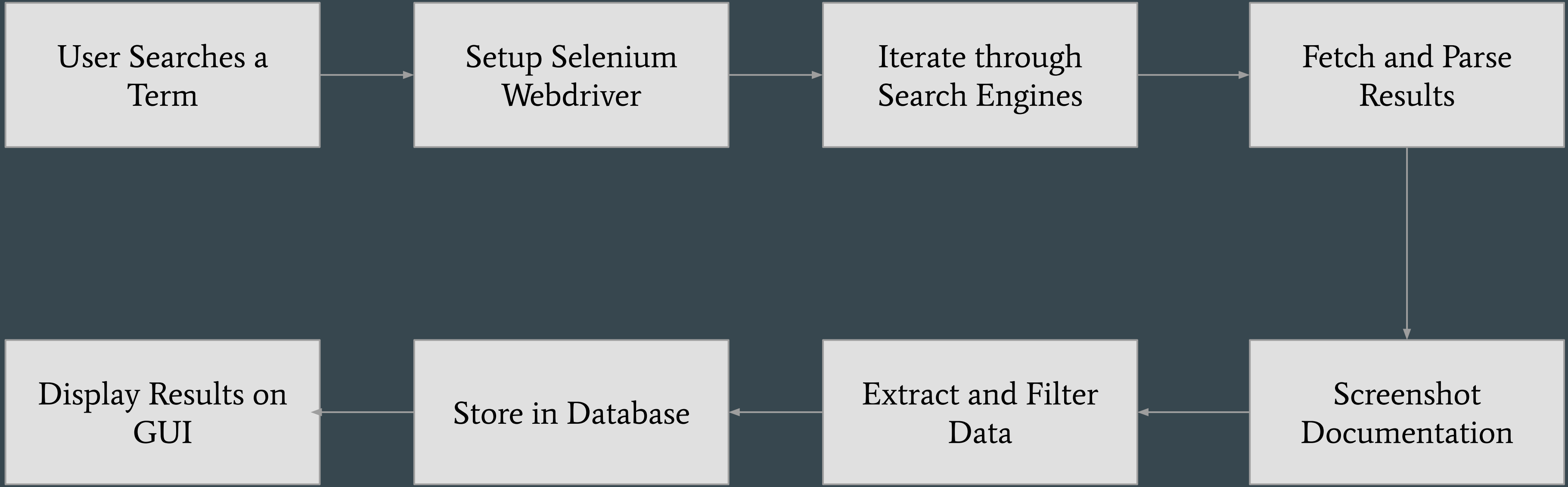
Table Result



```sql
-- Table used for the project
select * from searchresults order by Frequency DESC;
```

URL	Frequency	SearchTerm
https://health.usnews.com/best-hospitals/pediatric-rankings/cancer	33	Childhood cancer treatment best hospitals USA
https://www.stjude.org/media-resources/news-releases/2023-medicine-scien...	33	Childhood cancer treatment best hospitals USA
https://www.beckershospitalreview.com/oncology/20-best-hospitals-for-pedi...	28	Childhood cancer treatment best hospitals USA
https://www.stjude.org/about-st-jude/honors-and-awards/best-childrens-ho...	23	Childhood cancer treatment best hospitals USA
https://cancercenters.cancer.gov/cancer-centers/st-jude-childrens-research-...	22	Childhood cancer treatment best hospitals USA
https://www.beckershospitalreview.com/oncology/5-best-children-s-hospitals...	17	Childhood cancer treatment best hospitals USA
https://www.cancer.gov/research/infrastructure/cancer-centers/find/stjude	17	Childhood cancer treatment best hospitals USA
https://www.hopkinsmedicine.org/kimmel-cancer-center/cancers-we-treat/pe...	17	Childhood cancer treatment best hospitals USA
https://www.chop.edu/centers-programs/cancer-center	15	Childhood cancer treatment best hospitals USA

Implementation Workflow



User Searches a Term

Setup Selenium Webdriver

Iterate through Search Engines

Fetch and Parse Results

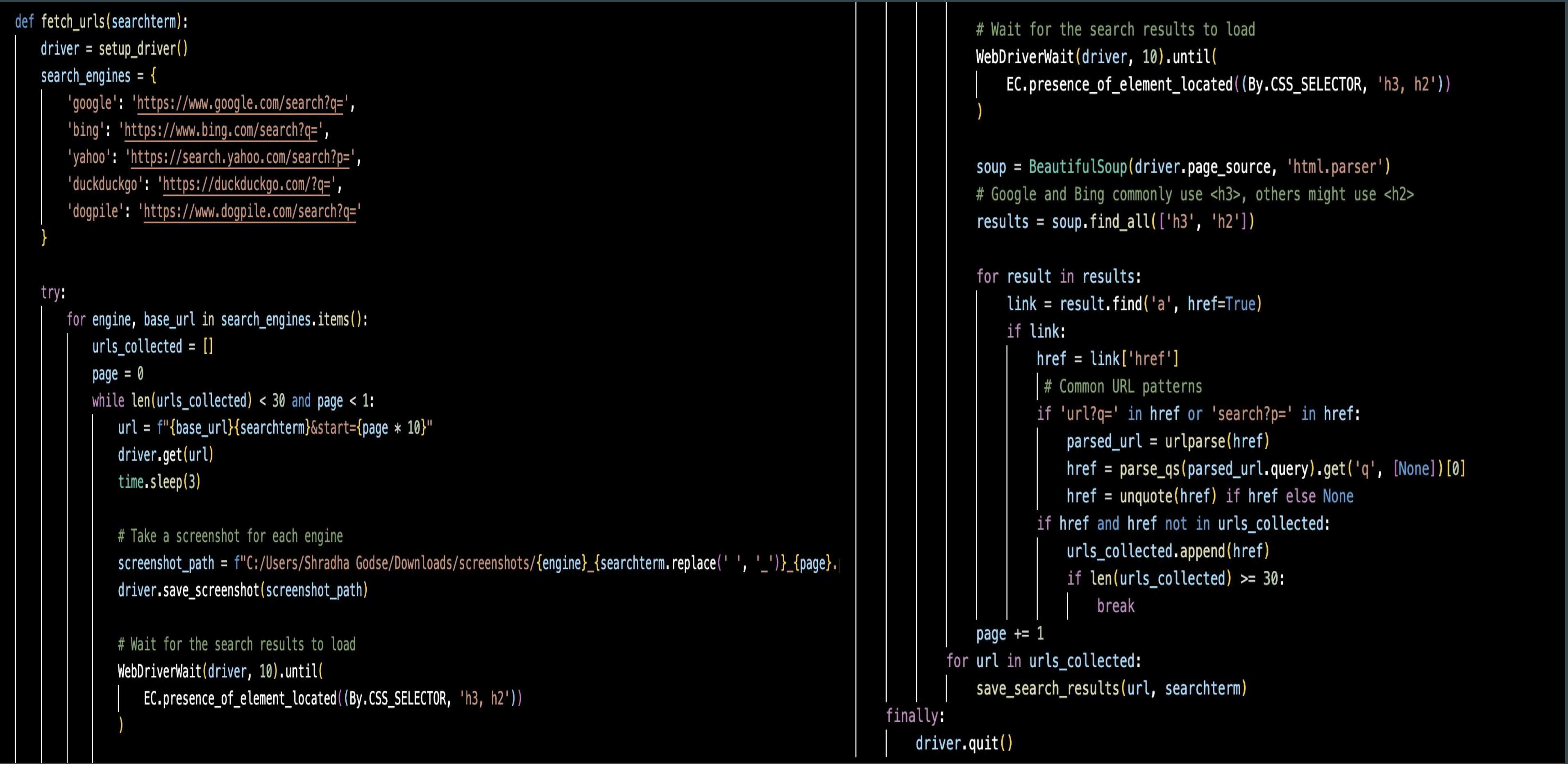
Screenshot Documentation

Extract and Filter Data

Store in Database

Display Results on GUI

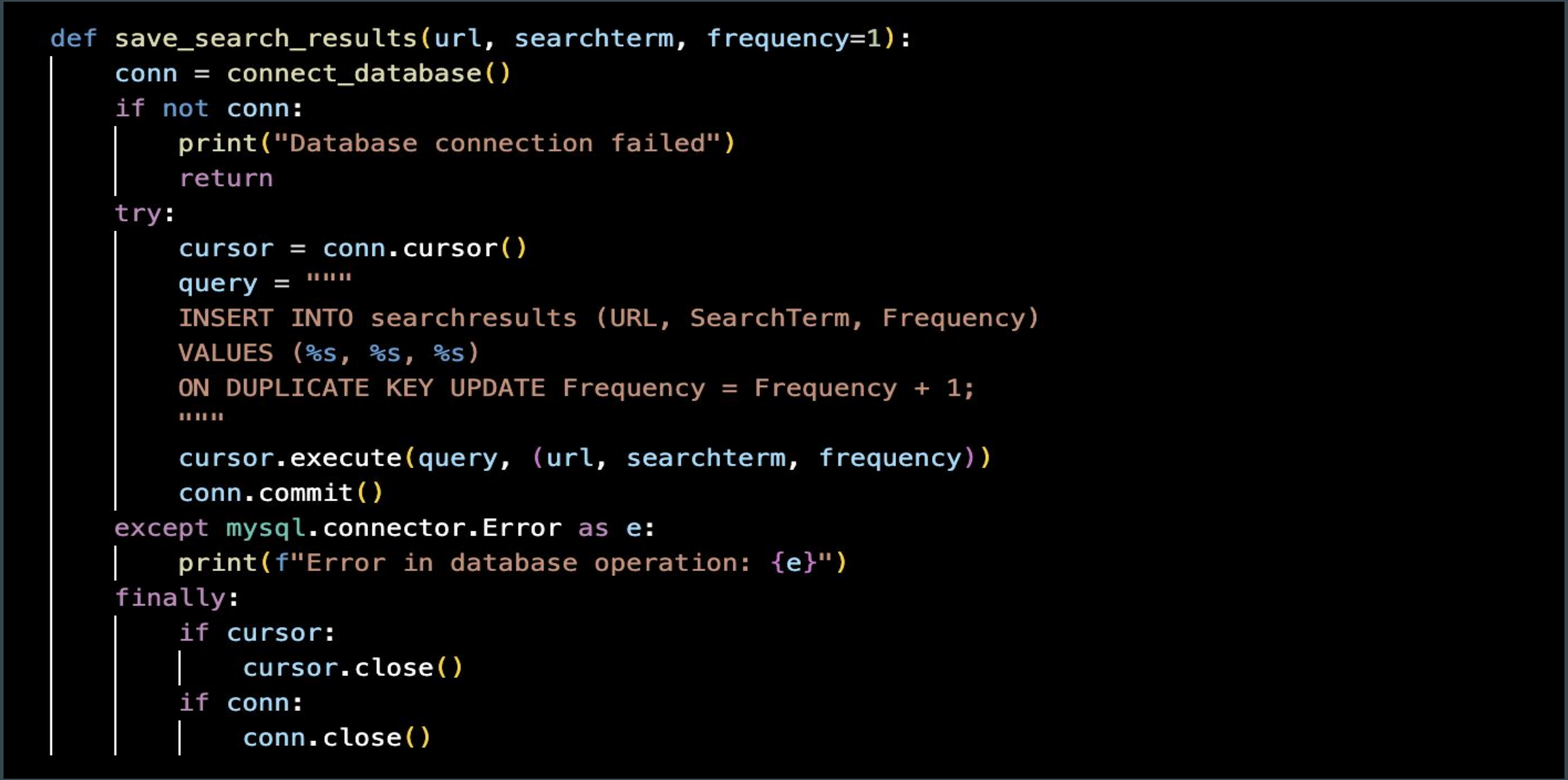
Fetch URL function



```
def fetch_urls(searchterm):
    driver = setup_driver()
    search_engines = {
        'google': 'https://www.google.com/search?q=',
        'bing': 'https://www.bing.com/search?q=',
        'yahoo': 'https://search.yahoo.com/search?p=',
        'duckduckgo': 'https://duckduckgo.com/?q=',
        'dogpile': 'https://www.dogpile.com/search?q='
    }

    try:
        for engine, base_url in search_engines.items():
            urls_collected = []
            page = 0
            while len(urls_collected) < 30 and page < 1:
                url = f"{base_url}{searchterm}&start={page * 10}"
                driver.get(url)
                time.sleep(3)

                # Take a screenshot for each engine
                screenshot_path = f"C:/Users/Shradha Godse/Downloads/screenshots/{engine}_{searchterm.replace(' ', '_')}_{page}.
                driver.save_screenshot(screenshot_path)

                # Wait for the search results to load
                WebDriverWait(driver, 10).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR, 'h3, h2'))
                )
```

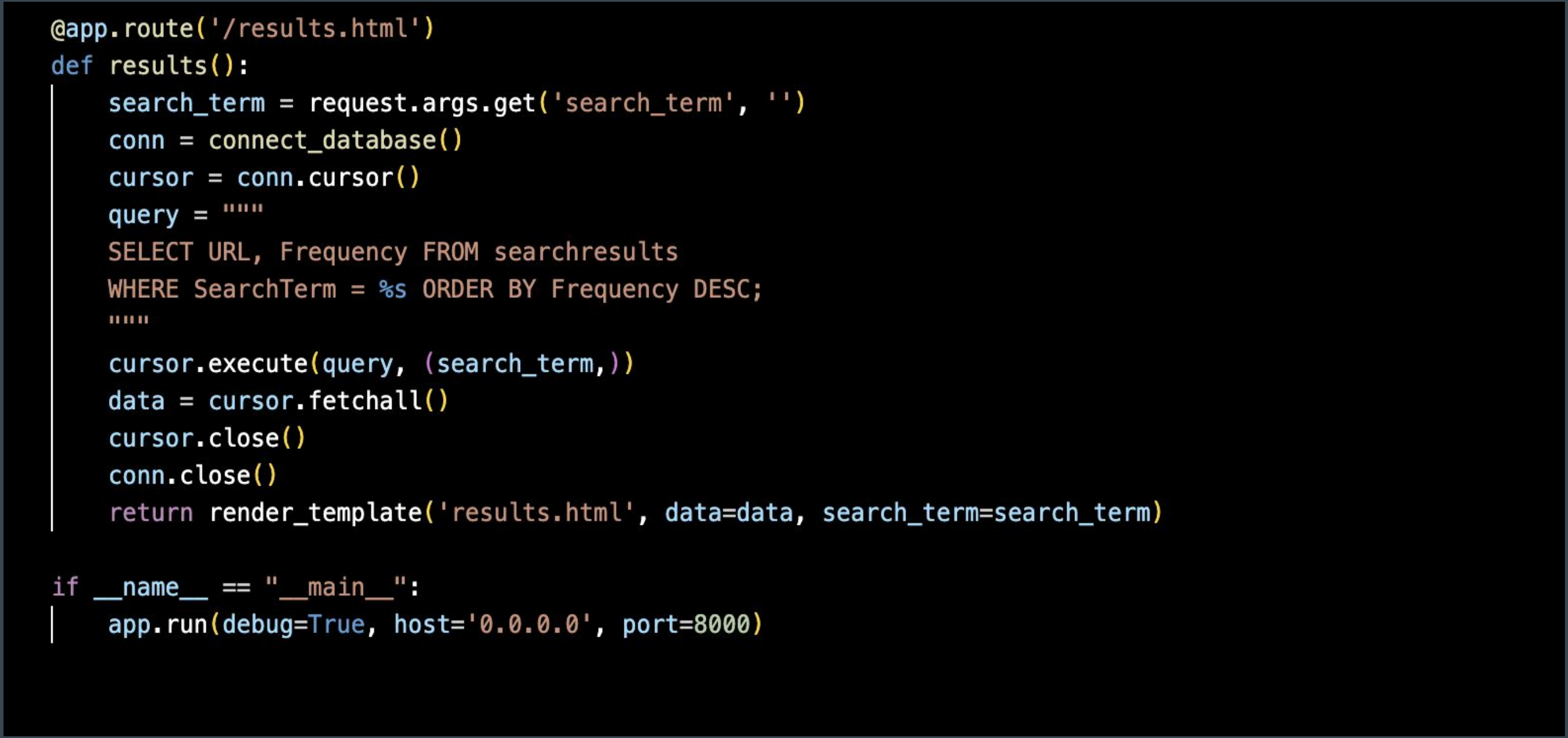
```
                # Wait for the search results to load
                WebDriverWait(driver, 10).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR, 'h3, h2'))
                )

                soup = BeautifulSoup(driver.page_source, 'html.parser')
                # Google and Bing commonly use <h3>, others might use <h2>
                results = soup.find_all(['h3', 'h2'])

                for result in results:
                    link = result.find('a', href=True)
                    if link:
                        href = link['href']
                        # Common URL patterns
                        if 'url?q=' in href or 'search?p=' in href:
                            parsed_url = urlparse(href)
                            href = parse_qs(parsed_url.query).get('q', [None])[0]
                            href = unquote(href) if href else None
                        if href and href not in urls_collected:
                            urls_collected.append(href)
                            if len(urls_collected) >= 30:
                                break
                page += 1
            for url in urls_collected:
                save_search_results(url, searchterm)
    finally:
        driver.quit()
```

Saving Search Results



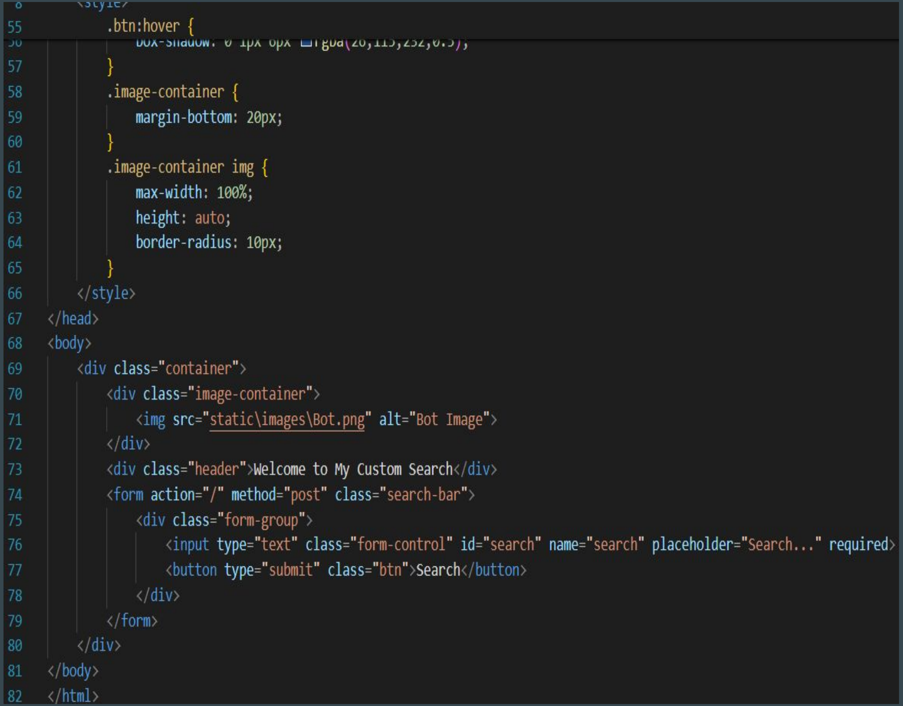
```python
def save_search_results(url, searchterm, frequency=1):
    conn = connect_database()
    if not conn:
        print("Database connection failed")
        return
    try:
        cursor = conn.cursor()
        query = """
        INSERT INTO searchresults (URL, SearchTerm, Frequency)
        VALUES (%s, %s, %s)
        ON DUPLICATE KEY UPDATE Frequency = Frequency + 1;
        """
        cursor.execute(query, (url, searchterm, frequency))
        conn.commit()
    except mysql.connector.Error as e:
        print(f"Error in database operation: {e}")
    finally:
        if cursor:
            cursor.close()
        if conn:
            conn.close()
```

Display Search Results



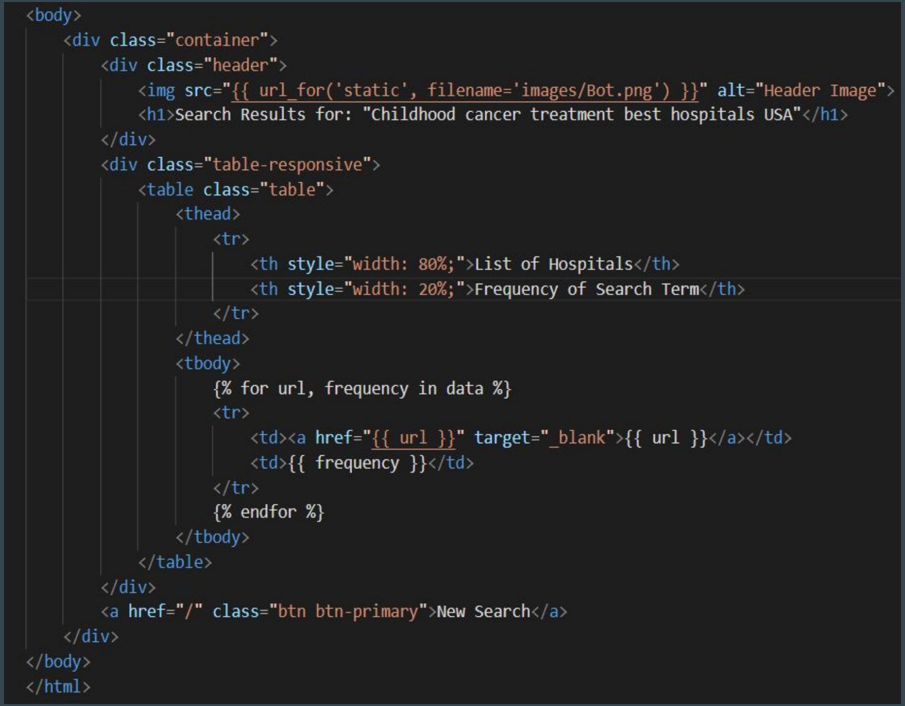
```
@app.route('/results.html')
def results():
    search_term = request.args.get('search_term', '')
    conn = connect_database()
    cursor = conn.cursor()
    query = """
    SELECT URL, Frequency FROM searchresults
    WHERE SearchTerm = %s ORDER BY Frequency DESC;
    """
    cursor.execute(query, (search_term,))
    data = cursor.fetchall()
    cursor.close()
    conn.close()
    return render_template('results.html', data=data, search_term=search_term)

if __name__ == "__main__":
    app.run(debug=True, host='0.0.0.0', port=8000)
```

index.html



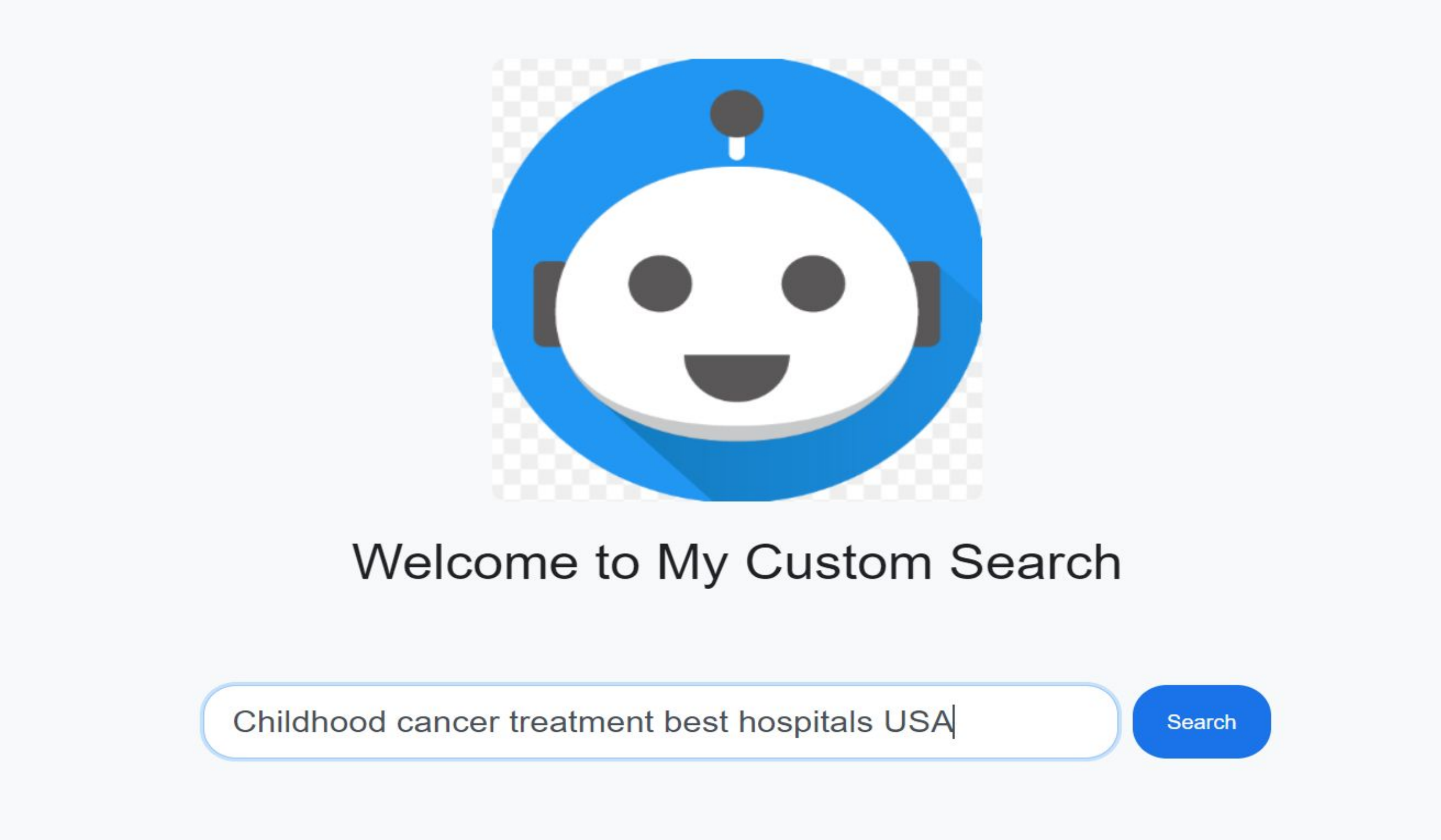
```
    <style>
        .btn:hover {
            box-shadow: 0 1px 6px rgba(26,115,232,0.5);
        }
        .image-container {
            margin-bottom: 20px;
        }
        .image-container img {
            max-width: 100%;
            height: auto;
            border-radius: 10px;
        }
    </style>
</head>
<body>
    <div class="container">
        <div class="image-container">
            <img src="static\images\Bot.png" alt="Bot Image">
        </div>
        <div class="header">Welcome to My Custom Search</div>
        <form action="/" method="post" class="search-bar">
            <div class="form-group">
                <input type="text" class="form-control" id="search" name="search" placeholder="Search..." required>
                <button type="submit" class="btn">Search</button>
            </div>
        </form>
    </div>
</body>
</html>
```

results.html



```
<body>
    <div class="container">
        <div class="header">
            <img src="{{ url_for('static', filename='images/Bot.png') }}" alt="Header Image">
            <h1>Search Results for: "Childhood cancer treatment best hospitals USA"</h1>
        </div>
        <div class="table-responsive">
            <table class="table">
                <thead>
                    <tr>
                        <th style="width: 80%;">List of Hospitals</th>
                        <th style="width: 20%;">Frequency of Search Term</th>
                    </tr>
                </thead>
                <tbody>
                    {% for url, frequency in data %}
                    <tr>
                        <td><a href="{{ url }}" target="_blank">{{ url }}</a></td>
                        <td>{{ frequency }}</td>
                    </tr>
                    {% endfor %}
                </tbody>
            </table>
        </div>
        <a href="/" class="btn btn-primary">New Search</a>
    </div>
</body>
</html>
```

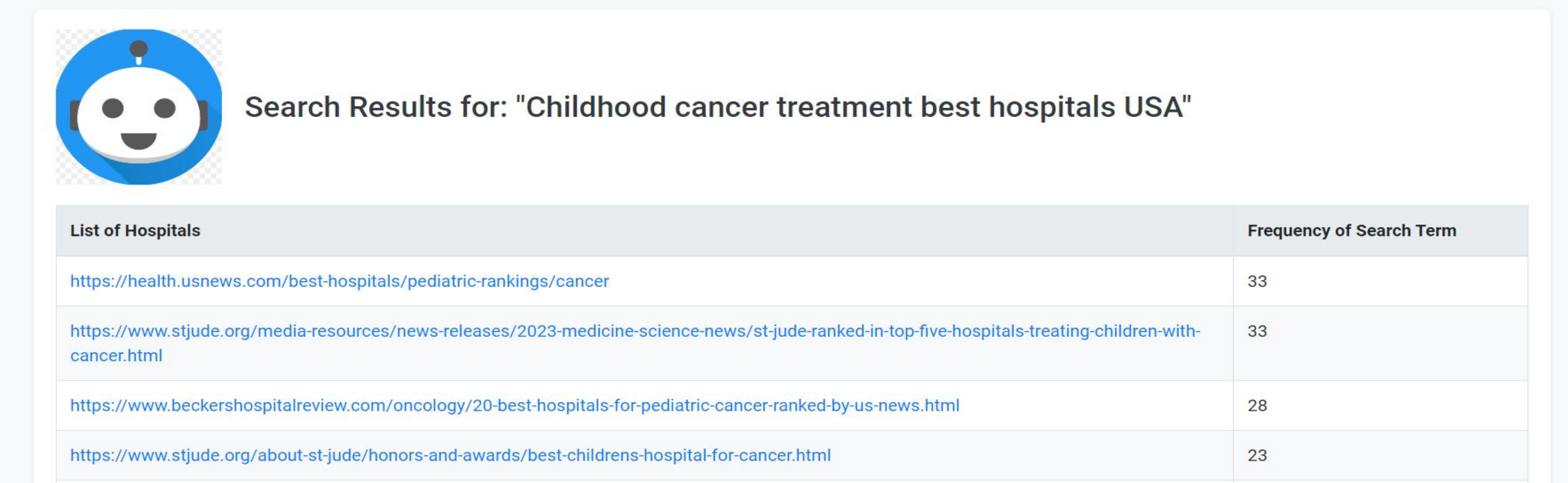
GUI Search Page



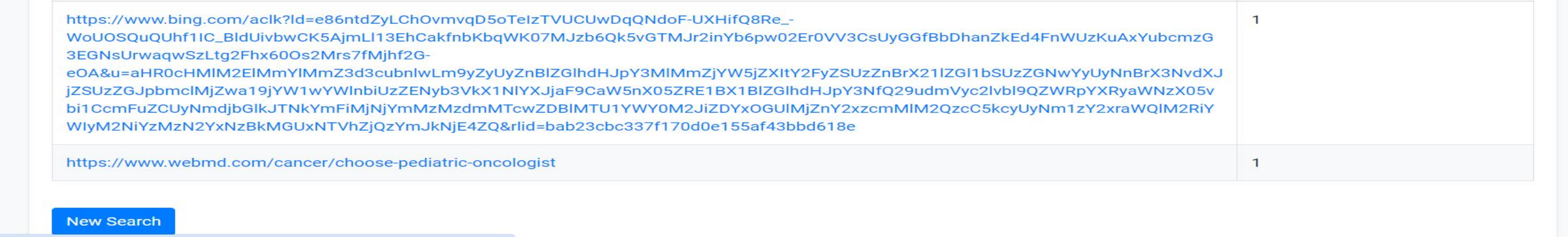
Welcome to My Custom Search

Childhood cancer treatment best hospitals USA

GUI Results Page



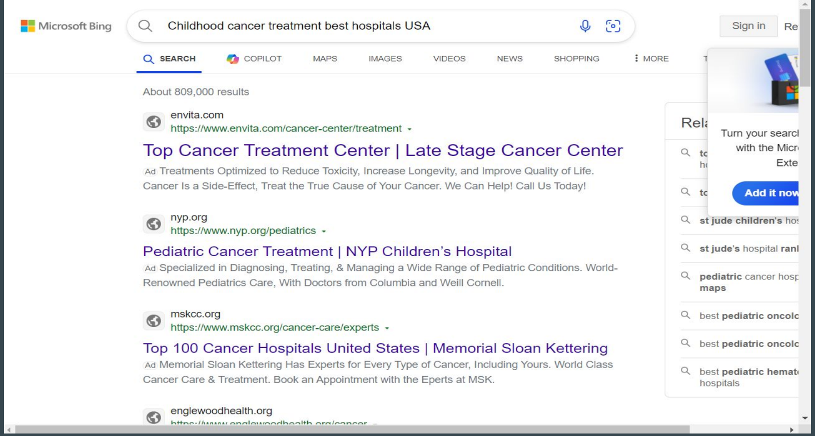
Search Results for: "Childhood cancer treatment best hospitals USA"



List of Hospitals	Frequency of Search Term
https://health.usnews.com/best-hospitals/pediatric-rankings/cancer	33
https://www.stjude.org/media-resources/news-releases/2023-medicine-science-news/st-jude-ranked-in-top-five-hospitals-treating-children-with-cancer.html	33
https://www.beckershospitalreview.com/oncology/20-best-hospitals-for-pediatric-cancer-ranked-by-us-news.html	28
https://www.stjude.org/about-st-jude/honors-and-awards/best-childrens-hospital-for-cancer.html	23
https://www.bing.com/aclk?ld=e86ntdZyLChOvmvqD5oTeIzTVUCUwDqQNdoF-UXHifQ8Re_-WoUOSQuQUhf1IC_BldUivbwCK5AjmLl13EhCakfnbKbqWK07MJzb6Qk5vGTMJr2inYb6pw02Er0VV3CsUyGGfBbDhanZkEd4FnWUzKuAxYubcmzG3EGNsUrwaqwSzLtg2Fhx60Os2Mrs7fMjhf2G-eOA&u=aHR0cHMlM2ElMmYlMmZ3d3cubnlwLm9yZyUyZnBlZGlhdHJpY3MlMmZjYW5jZXItY2FyZSUzZnBrX21lZGl1bSUzZGNwYyUyNnBrX3NvdXJjZSUzZGJpbmclMjZwa19jYW1wYWlnbiUzZENyb3VkX1NlYXJjaF9CaW5nX05ZRE1BX1BlZGlhdHJpY3NfQ29udmVyc2lvbl9QZWRpYXRyaWNzX05vbi1CcmFuZCUyNmdjbGlkJTNkYmFiMjNjYmMzMzdmMTcwZDBlMTU1YWY0M2JiZDYxOGUlMjZnY2xzcmMlM2QzcC5kcyUyNm1zY2xraWQlM2RiYWIyM2NiYzMzN2YxNzBkMGUxNTVhZjQzYmJkNjE4ZQ&rlid=bab23cbc337f170d0e155af43bbd618e	1
https://www.webmd.com/cancer/choose-pediatric-oncologist	1

New Search

Screenshot Saves using Selenium



Childhood cancer treatment best hospitals USA

About 809,000 results

envita.com
https://www.envita.com/cancer-center/treatment

Top Cancer Treatment Center | Late Stage Cancer Center

Ad Treatments Optimized to Reduce Toxicity, Increase Longevity, and Improve Quality of Life. Cancer Is a Side-Effect, Treat the True Cause of Your Cancer. We Can Help! Call Us Today!

nyp.org
https://www.nyp.org/pediatrics

Pediatric Cancer Treatment | NYP Children's Hospital

Ad Specialized in Diagnosing, Treating, & Managing a Wide Range of Pediatric Conditions. World-Renowned Pediatrics Care, With Doctors from Columbia and Weill Cornell.

mskcc.org
https://www.mskcc.org/cancer-care/experts

Top 100 Cancer Hospitals United States | Memorial Sloan Kettering

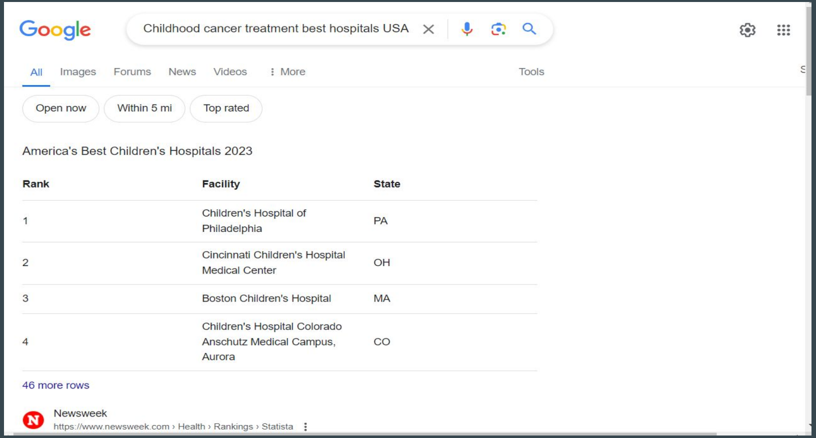
Ad Memorial Sloan Kettering Has Experts for Every Type of Cancer, Including Yours. World Class Cancer Care & Treatment. Book an Appointment with the Eperts at MSK.

englewoodhealth.org

Rela

Turn your searc
with the Micr
Exte

- st jude children's hos
- st jude's hospital ran
- pediatric cancer hosp maps
- best pediatric oncolo
- best pediatric oncolo
- best pediatric hemat hospitals



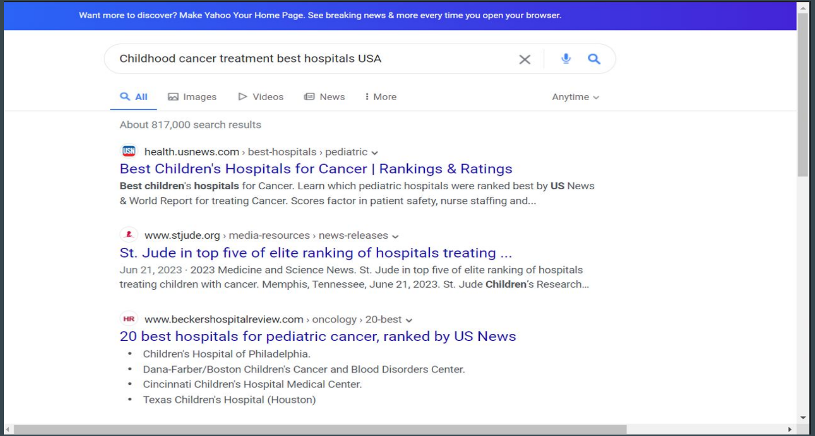
Childhood cancer treatment best hospitals USA

America's Best Children's Hospitals 2023

Rank	Facility	State
1	Children's Hospital of Philadelphia	PA
2	Cincinnati Children's Hospital Medical Center	OH
3	Boston Children's Hospital	MA
4	Children's Hospital Colorado Anschutz Medical Campus, Aurora	CO

46 more rows

Newsweek
https://www.newsweek.com › Health › Rankings › Statista



Want more to discover? Make Yahoo Your Home Page. See breaking news & more every time you open your browser.

Childhood cancer treatment best hospitals USA

About 817,000 search results

health.usnews.com › best-hospitals › pediatric

Best Children's Hospitals for Cancer | Rankings & Ratings

Best children's **hospitals** for Cancer. Learn which pediatric hospitals were ranked best by **US** News & World Report for treating Cancer. Scores factor in patient safety, nurse staffing and...

www.stjude.org › media-resources › news-releases

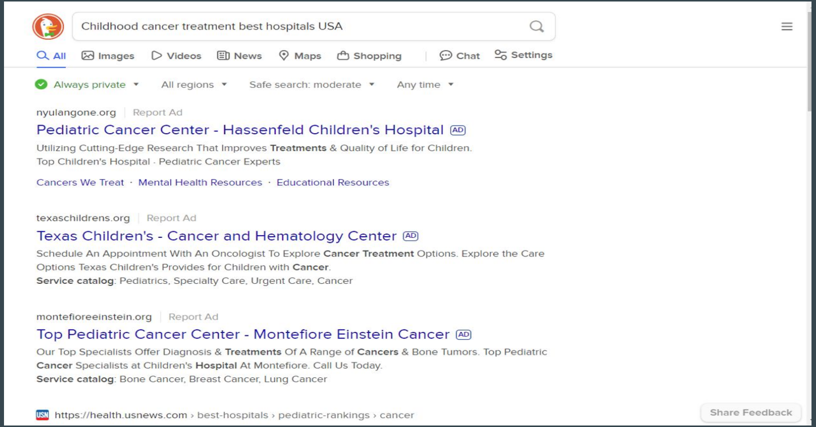
St. Jude in top five of elite ranking of hospitals treating ...

Jun 21, 2023 · 2023 Medicine and Science News. St. Jude in top five of elite ranking of hospitals treating children with cancer. Memphis, Tennessee, June 21, 2023. St. Jude **Children**'s Research...

www.beckershospitalreview.com › oncology › 20-best

20 best hospitals for pediatric cancer, ranked by US News

- Children's Hospital of Philadelphia.
- Dana-Farber/Boston Children's Cancer and Blood Disorders Center.
- Cincinnati Children's Hospital Medical Center.
- Texas Children's Hospital (Houston)



Childhood cancer treatment best hospitals USA

nyulangone.org | Report Ad

Pediatric Cancer Center - Hassenfeld Children's Hospital AD

Utilizing Cutting-Edge Research That Improves **Treatments** & Quality of Life for Children.
Top Children's Hospital · Pediatric Cancer Experts

Cancers We Treat · Mental Health Resources · Educational Resources

texaschildrens.org | Report Ad

Texas Children's - Cancer and Hematology Center AD

Schedule An Appointment With An Oncologist To Explore **Cancer Treatment** Options. Explore the Care Options Texas Children's Provides for Children with **Cancer**.
Service catalog: Pediatrics, Specialty Care, Urgent Care, Cancer

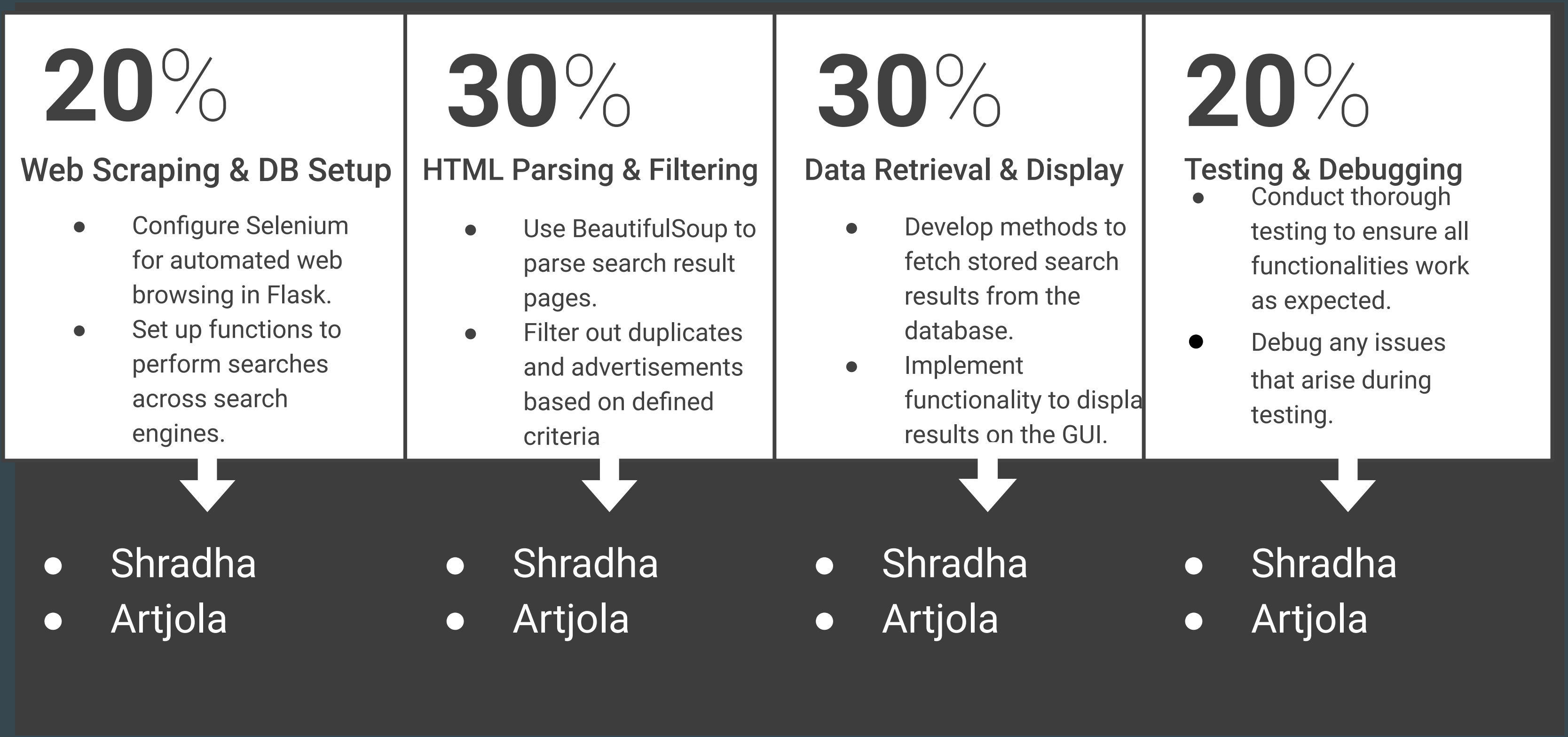
montefioreeinstein.org | Report Ad

Top Pediatric Cancer Center - Montefiore Einstein Cancer AD

Our Top Specialists Offer Diagnosis & **Treatments** Of A Range of **Cancers** & Bone Tumors. Top Pediatric **Cancer** Specialists at Children's **Hospital** At Montefiore. Call Us Today.
Service catalog: Bone Cancer, Breast Cancer, Lung Cancer

https://health.usnews.com › best-hospitals › pediatric-rankings › cancer

Work Division



20%

Web Scraping & DB Setup

- Configure Selenium for automated web browsing in Flask.
- Set up functions to perform searches across search engines.

- Shradha
- Artjola

30%

HTML Parsing & Filtering

- Use BeautifulSoup to parse search result pages.
- Filter out duplicates and advertisements based on defined criteria

- Shradha
- Artjola

30%

Data Retrieval & Display

- Develop methods to fetch stored search results from the database.
- Implement functionality to displa results on the GUI.

- Shradha
- Artjola

20%

Testing & Debugging

- Conduct thorough testing to ensure all functionalities work as expected.
- Debug any issues that arise during testing.

- Shradha
- Artjola

Summary

- Efficiently fetches relevant URLs while avoiding duplicates and ads.
- Ensures high-quality search results for user queries.

Benefits:

- Combines database constraints and HTML parsing for effective filtering.
- Delivers clean and relevant search results based on the count of frequency search terms on the GUI Browser.

Demo