Medline Search Engine

- Harnessing Big Data & Analytics for Medical Information Retrieval
- CIS 612 - Big Data and Para Database Systems
- Guided by – Sunnie S. Chung, Ph.D.

MEET OUR TEAM

Sushma Avala - 2885387

Dinky Mishra – 2864923

Aditya Sairam Pullabhatla - 2863159

Alim Khan Abdul - 2882808

Introduction:

- Goal : Develop a scalable and intelligent search engine for medical information, leveraging information retrieval techniques learnt from class.

- Objective : Implement advanced NLP techniques (e.g., TF-IDF, Cosine Similarity) for precise search matching, and create a user-friendly platform for real-time medical data retrieval.

Tools and Technologies:

1. **Frontend Development -** Next.js
2. **Backend Development -** Next.js
3. **Database Management -** MongoDB, MySQL
4. **Web Scraping -** Puppeteer
5. **Natural Language Processing (NLP) -** Pluralize, Lemmatizer, Stem-Porter
6. **Performance Optimization-** Batched database operations

Knowledge Base and Database Design

Optimized for Efficient Data Management

MongoDB for Raw Data ● Facilitates storage of unprocessed HTML and metadata for backup and initial processing.	MySQL for Structured Data ● Hosts cleaned articles, term Frequencies and Inverted Index for quick retrieval.
Inverted Index Optimization ● Enhances search efficiency by mapping terms to relevant documents with TF-IDF scores.	Schema Flexibility ● Designed to scale and accommodate expanding datasets seamlessly.

Data Collection Process

- Gathering Raw Medical Data
- Automates Scraping
- Node.js and Puppeteer used to scrape 4500 articles.
- Raw Data Storage
- Articles stored as HTML in MongoDB for reliability.
- Scalable and Repeatable
- System designed for efficient and repeatable scraping

Data Validation

- Ensure Data Integrity :
- **Data Validation** : Cross-referenced scraped articles with source counts.
- **Duplicate Removal** : Ensured unique entries by identifying and removing duplicates.
- **Content Accuracy** :Verifies and cleaned content for relevance and consistency.

HTML Content Cleaning

Eliminated unnecessary HTML elements (e.g., <script>, <style>).

Ad and Navigation Removal: Filtered out irrelevant content such as ads and links.

Identified xPaths containing medically relevant content and extracted all the necessary information.

Tokenization and Lemmatization

- Breaking Down Cleaned Data
- **Tokenization:** Split text into individual tokens (words) for analysis.
- **Lemmatization:** Converted words to their base forms for consistency.
- **Stemming Comparison:** Preferred over stemming to retain grammatical meaning.

Dataset Overview

Articles Scraped: ~4,500 medical articles from MedlinePlus Encyclopedia.

Terms Processed: Over 1.4 million terms indexed after preprocessing.

Comprehensive Coverage: Covers diseases, symptoms, treatments, and medications.

Ranking Algorithms

Algorithms	Purpose	Use Case
1.NLP techniques	Tokenization , Stemming , Lemmatization	Used for preprocessing abd normalizing the text .
2.TF-IDF (Vectorization)	Identifies important terms in a document within the dataset.	Finds terms most relevant the user's query.
3.Cosine Similarity	Measures Similarity between query and document Vectors	Ranks documents by comparing vector angles.

Inverted Index

Inverted Index from scraped data.	Dynamic Query Inverted Index
• A static inverted index built from Medline's 4,500 articles. • Stores terms, term frequencies (TF), and document frequencies.	• Generated during user query processing. • Captures query term frequencies and builds a temporary index for matching.

Dictionary Table

Mapping Terms with Metadata

High-Level Indexing: Stores unique terms and their metadata for efficient lookups.

Fields in Dictionary Table: Term, Document Frequency (DF), Collection Frequency.

Purpose: Acts as an entry point for term-to-document mappings.

Posting Table

Storing Document-Level Term Details

Detailed Mappings: Maps terms to document IDs and metadata like term frequency (TF).

Fields in Posting Table: Document ID (DOC_ID), Term Frequency (TF), and optional positions.

Purpose: Facilitates document-level relevance computation and query matching.

Cosine Similarity

Query and Document Vectors: Built using TF-IDF weights for terms.

Similarity Calculation: Measures angle between query and document vectors.

Ranking Results: Scores determine the most relevant documents.

Implementation Workflow

User Interface (Next.js Frontend) → Backend Server (Next.js built-in routing system → Inverted Index (MySQL) → Similarity Sources (Cosine Similarity Calculation) → Frontend UI Display Results

Challenges and Solutions

Challenges :

XPath Issues During Web Scraping:

- Problem: Initially used XPaths failed to locate all content due to structural changes on the MedlinePlus website.
- Cause: Content was located in additional, previously unconsidered classes.

Inefficient Database Query Execution:

- Problem: Querying ~4,000 documents individually for each user query led to a response time of 60–75 minutes.
- Cause: Lack of a consolidated query mechanism.

Resolutions :

Updated XPath Selectors:

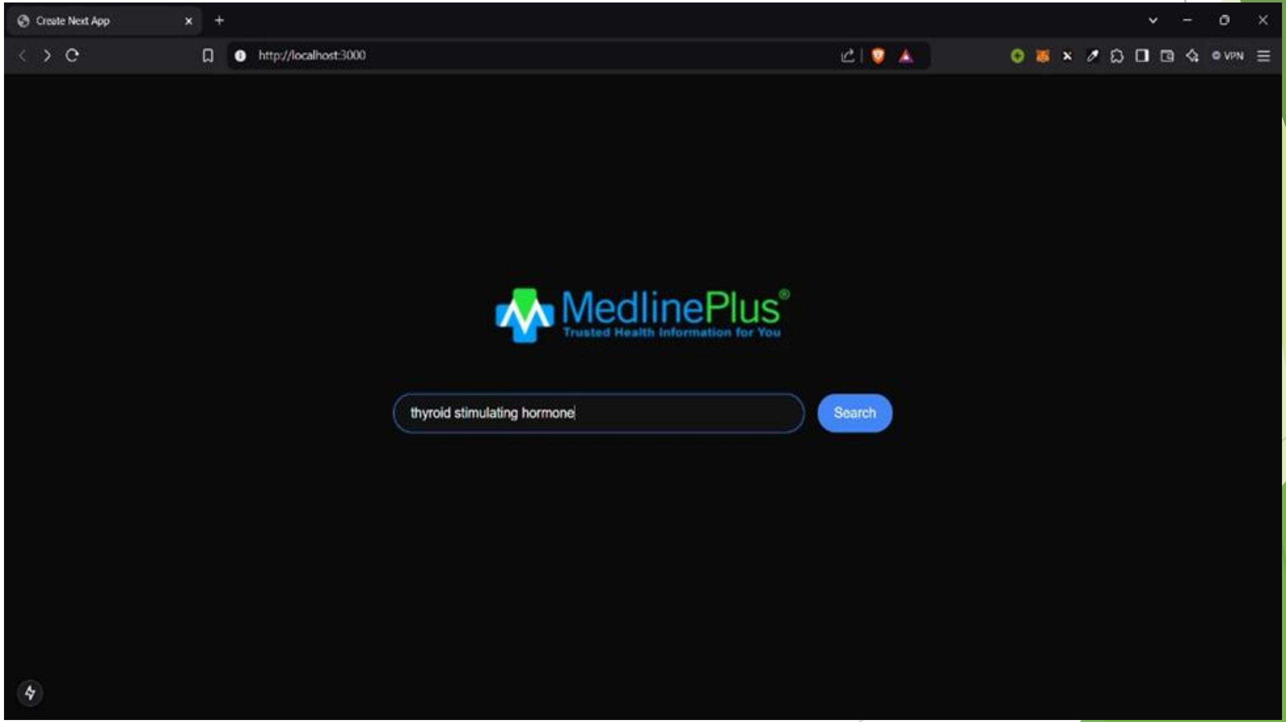
- Solution: Debugged and identified the correct classes, updating XPaths to capture all relevant data consistently.

Optimized Query Processing:

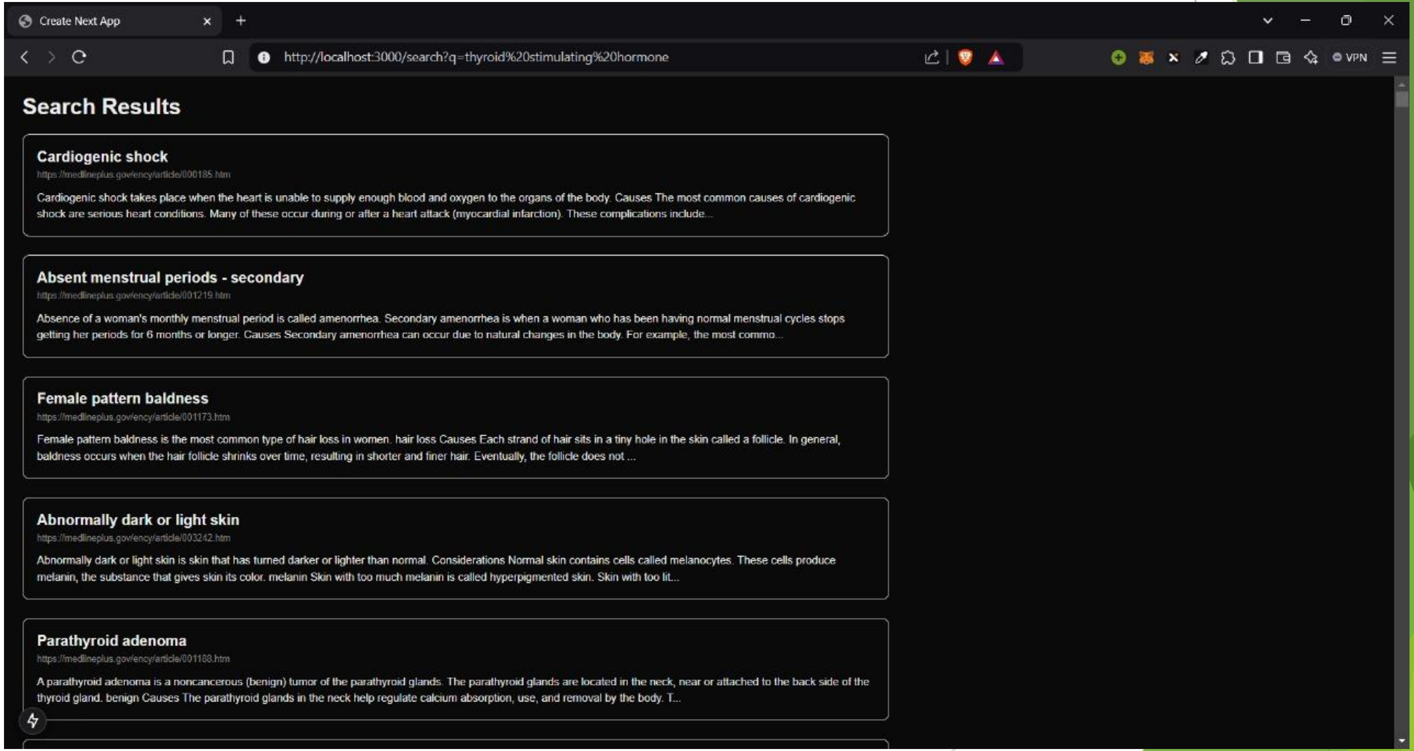
- Solution: Introduced consolidated queries, retrieving all relevant data in a single database call.
- Impact: Significantly reduced response time, enabling near real-time query results.

Results

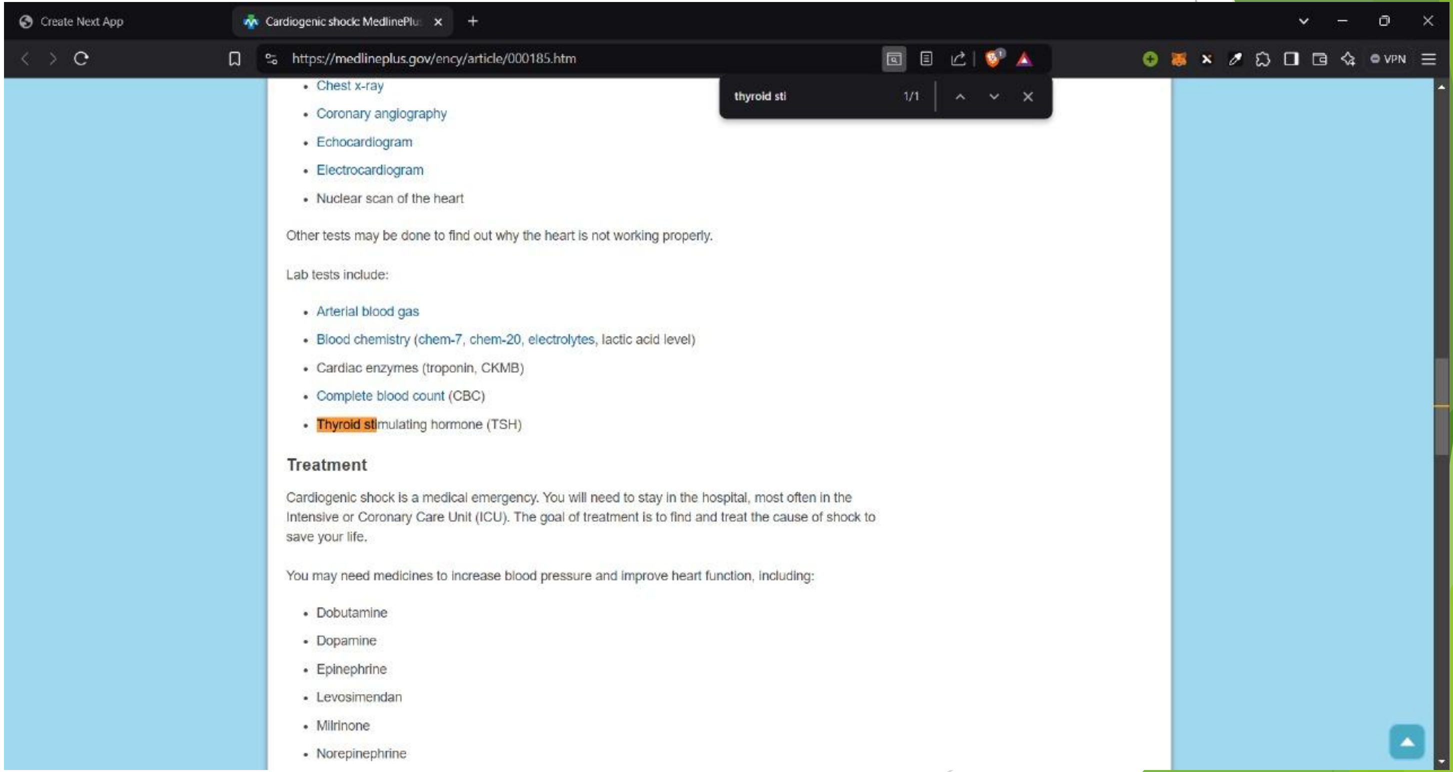
Performance Metrics with a User-Centric Interface



Results


Create Next App
http://localhost:3000/search?q=thyroid%20stimulating%20hormone
VPN
Search Results
Cardiogenic shock
https://medlineplus.gov/ency/article/000185.htm
Cardiogenic shock takes place when the heart is unable to supply enough blood and oxygen to the organs of the body. Causes The most common causes of cardiogenic shock are serious heart conditions. Many of these occur during or after a heart attack (myocardial infarction). These complications include...
Absent menstrual periods - secondary
https://medlineplus.gov/ency/article/001219.htm
Absence of a woman's monthly menstrual period is called amenorrhea. Secondary amenorrhea is when a woman who has been having normal menstrual cycles stops getting her periods for 6 months or longer. Causes Secondary amenorrhea can occur due to natural changes in the body. For example, the most commo...
Female pattern baldness
https://medlineplus.gov/ency/article/001173.htm
Female pattern baldness is the most common type of hair loss in women. hair loss Causes Each strand of hair sits in a tiny hole in the skin called a follicle. In general, baldness occurs when the hair follicle shrinks over time, resulting in shorter and finer hair. Eventually, the follicle does not ...
Abnormally dark or light skin
https://medlineplus.gov/ency/article/003242.htm
Abnormally dark or light skin is skin that has turned darker or lighter than normal. Considerations Normal skin contains cells called melanocytes. These cells produce melanin, the substance that gives skin its color. melanin Skin with too much melanin is called hyperpigmented skin. Skin with too lit...
Parathyroid adenoma
https://medlineplus.gov/ency/article/001188.htm
A parathyroid adenoma is a noncancerous (benign) tumor of the parathyroid glands. The parathyroid glands are located in the neck, near or attached to the back side of the thyroid gland. benign Causes The parathyroid glands in the neck help regulate calcium absorption, use, and removal by the body. T...

Results



Create Next App

Cardiogenic shock: MedlinePlu

https://medlineplus.gov/ency/article/000185.htm

thyroid sti 1/1

- Chest x-ray
- Coronary angiography
- Echocardiogram
- Electrocardiogram
- Nuclear scan of the heart

Other tests may be done to find out why the heart is not working properly.

Lab tests include:

- Arterial blood gas
- Blood chemistry (chem-7, chem-20, electrolytes, lactic acid level)
- Cardiac enzymes (troponin, CKMB)
- Complete blood count (CBC)
- Thyroid stimulating hormone (TSH)

Treatment

Cardiogenic shock is a medical emergency. You will need to stay in the hospital, most often in the Intensive or Coronary Care Unit (ICU). The goal of treatment is to find and treat the cause of shock to save your life.

You may need medicines to increase blood pressure and improve heart function, including:

- Dobutamine
- Dopamine
- Epinephrine
- Levosimendan
- Milrinone
- Norepinephrine

Search Speed: 3-6 sec approx. for each search

Scalability: Designed to handle larger datasets with dynamic query indexing.

Conclusion

Key Takeaways

Project Achievements: Indexed ~4,500 articles with real-time search functionality.	**Impact on Accessibility:** Improved access to structured medical data with a user-friendly interface.
Future Directions: Potential for multilingual support and semantic search.	**Final Remark:** Showcased AI's potential in transforming medical data retrieval.

Any Questions ?

Thank you