PROJECT ON INFORMATION RETRIVAL (CSP – 429)

SEARCH ENGINE DEVELOPMENT by

Sruthi Kondapalli(A20554780)

UNDER THE GUIDANCE OF

Mr. Jawahar Panchal
Professor (CSP-429)
Department of Computer Science

Abstract:

This project entails the development of a search engine leveraging various technologies such as web crawling, text indexing, and query processing. The objectives included creating a scalable solution capable of indexing web documents, enabling efficient search capabilities, and providing relevant results to user queries. The next steps involve potential enhancements such as improving search accuracy, handling larger datasets, and enhancing user experience.

Overview:

The search engine project aims to address the need for efficient information retrieval from web documents. Leveraging Scrapy for web crawling, Scikit-Learn for text indexing, and Flask for query processing, the solution provides a comprehensive approach to search functionality. Relevant literature on web crawling, text indexing, and search algorithms informed the design and implementation of the proposed system.

Design:

The system encompasses three main components: the web crawler, the indexer, and the query processor. The web crawler, developed using Scrapy, retrieves web documents in HTML format from specified domains. The indexer, based on Scikit-Learn, constructs an inverted index using TF-IDF representation and cosine similarity scoring. The query processor, implemented with Flask, handles user queries, validates input, and returns top-ranked search results.

Architecture:

The software components interact as follows:
1.The web crawler initiates crawling from seed URLs, downloading HTML documents.
2.Extracted content is indexed using TF-IDF representation and cosine similarity scoring.
3.The inverted index is stored in pickle format for efficient retrieval.
4.The Flask-based query processor accepts user queries, processes them against the inverted index, and returns relevant search results.

Operation:

To operate the search engine:
1.Run the web crawler using Scrapy to obtain web documents.
2.Execute the indexer to construct the inverted index.
3.Start the Flask server for query processing.
4.Access the search interface via a web browser and input queries.

Conclusion:

The search engine project demonstrates success in providing a functional system for information retrieval. Results indicate efficient indexing and retrieval of web documents. However, further optimization may be necessary for handling larger datasets and improving search accuracy. Caution is advised regarding scalability and resource utilization in production environments.

Data Sources:

Web documents were obtained from quotes.toscrape.com Additional datasets may be integrated for broader search capabilities.

Test Cases:

Test cases were employed to validate system functionality, covering web crawling, indexing, query processing, and result retrieval. Further testing frameworks and coverage analysis may be implemented for comprehensive validation.

Source Code:

The source code for the project is provided in Python files, including quote.py for the web crawler, indexer.py for the indexer, and app.py for the query processor. Dependencies include Python 3.10+, Scrapy 2.11+, Scikit-Learn 1.2+, and Flask 2.2+.

CODEBASE: https://github.com/SruthiKondapalli/Information-RetrievalProject

REFERENC_WEBSITE: "https://quotes.toscrape.com/"

OUTPUTS:

Search Engine

quotes Search

Search Results

https://quotes.toscrape.com/page/10/ - Score: 0

https://quotes.toscrape.com/page/4/ - Score: 0

https://quotes.toscrape.com/page/3/ - Score: 0

https://quotes.toscrape.com/page/3/ - Score: 0

https://quotes.toscrape.com/page/3/ - Score: 0

https://quotes.toscrape.com/page/3/ - Score: 0

https://quotes.toscrape.com/page/4/ - Score: 0

https://quotes.toscrape.com/page/4/ - Score: 0

https://quotes.toscrape.com/page/4/ - Score: 0

https://quotes.toscrape.com/page/4/ - Score: 0

Quotes to Scrape

Login

"The truth." Dumbledore sighed. "It is a beautiful and terrible thing, and should therefore be treated with great caution."
by J.K. Rowling (about)
Tags: truth

"I'm the one that's got to die when it's time for me to die, so let me live my life the way I want to."
by Jimi Hendrix (about)
Tags: death life

"To die will be an awfully big adventure."
by J.M. Barrie (about)
Tags: adventure love

"It takes courage to grow up and become who you really are."
by E.E. Cummings (about)
Tags: courage

Top Ten tags

love
inspirational
life
humor
books
reading
friendship
friends
truth
smile

Bibliography:

1.Bird, Steven, Edward Loper, and Ewan Klein. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., 2009.
2.Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
3.Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
4.Ronacher, A. (2022). Flask Documentation. Retrieved from https://flask.palletsprojects.com/.