Tiny Search Engine

Justin Garrigus, Miguel Hernandez,
Pascal Rioux-Coulombe, Philip Rhodes

Github

https://github.com/justinmgarrigus/Tiny-Search-Engine

Motivation

A search engine is a type of program designed to provide a list of documents given a search query. It presents a number of computational challenges across a diverse array of problems including web crawling, indexing, and search query intent classification. This project hopes to explore these areas by creating a tiny search engine that allows users to search for a desired topic and receive an indexed, ranked list of hyperlinks that match this search.

Significance

Search engines have provided the backbone for web navigation since their standardization in the 1990s and continue to maintain a large market share on software development efforts. Be they the indexing force for generalized hyperlinks to web pages in the cases of Google, Bing, and Yahoo, or for videos in the cases of Youtube and Twitch, search engines provide unique challenges for natural language processing from the creation and maintenance of crawlers that gather information from websites to the semantic analysis that ranks and categorizes this information.

Search engines will continue to be an integral part of web interfaces for the foreseeable future, therefore this project provides its members with a foundational insight into such an important field application of NLP.

Objectives

The output of the project should be a single program—written in several different languages including Python, SQL, and C++—that at least supports the following high-level operations:

- Scraping keywords from a given website.
- Indexing a hyperlink in a database alongside its classifiers.
- Allowing the user to input a search query and optionally autocompleting their query.
- Retrieving a ranked list of hyperlinks that match the query.

The goal of this project is not to design a perfect search engine or a program that is usable in everyday life; instead, it serves as an opportunity to learn more about the domain and combining natural language processing with big data.

```
user:/tinysearchengine$ ./tinysearchengine -i https://en.wikipedia.org/wiki/Natural_language_processing --verbose
Indexing "https://en.wikipedia.org/wiki/Natural_language_processing"... Done
  Keywords:
    natural language processing
    speech recognition
    artificial intelligence
    nlp
    machine translation
    [45 others...]
  Inserting into database_hyperlinks... Done
    New size: 1567 links

user:/tinysearchengine$ ./tinysearchengine
Insert a query: natural language?
Suggestions:
  1.) natural language processing
  2.) natural language understanding
  3.) languages that are natural
  4.) natural language processing machine learning
  5.) what are natural languages
Insert a query or select an item: 1
Links:
  1.) Natural Language Processing - Wikipedia
        Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence
        concerned with the interactions between ...
  2.) What is Natural Language Processing? - IBM
        Natural language processing (NLP) refers to the branch of computer science—and more specifically, the
        branch of artificial intelligence or AI— ...
  3.) Natural Language Processing (NLP): What it is and why it matters - SAS
        Natural language processing helps computers communicate with humans in their own language and scales
        other language-related tasks. For example, NLP makes it ...
  4.) What is Natural Language Processing? An Introduction to NLP
        Natural language processing (NLP) is the ability of a computer program to understand human language as it is
        spoken and written -- referred to as natural ...
  5.) What is Natural Language Processing? Introduction to NLP
        Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand,
        and derive meaning from human language in a ...
Insert a query or select an item: 1
  Opening "https://en.wikipedia.org/wiki/Natural_language_processing"...
```

An example execution of the program.

Features

At a minimum, the search engine should support these features:

- Indexing of a website provided via user input. The user should be able to provide a hyperlink to a website (for instance, “https://en.wikipedia.org/wiki/Natural_language_processing”) or a text file containing a list of websites, to which the program would:
  - Read the plain text from the website.
  - Associate the link with a collection of keywords and classifiers.
  - Store the link and classifiers into a database with a language like SQL.
- Allow the user to input a search query in plain english.
- Offer related link suggestions based on user searches.

    - Probabilistically for autocompletions of searches.
    - Based on semantic similarity for historical searches.
    - Based on the content of documents previously indexed.
- Retrieve a list of websites that the query may apply to.
    - Websites should be ranked by their relevance based on their content, the number of other websites that reference them, and more.

More features can be added to the engine to increase its usability, including:

- Mass re-indexing over time (websites are indexed more than one time to increase the relevance of provided search terms).
- Employing machine learning to improve website ranking and query autocomplete (dynamically training a model based on the type of information a user provides).
- Exact keyword searches (queries that are surrounded in double-quotes will yield documents that contain exactly the given string of text).

References

[1] Britannica, T. Editors of Encyclopaedia. "search engine." Encyclopedia Britannica, January 12, 2022. Retrieved https://www.britannica.com/technology/search-engine.