Viet Nam National University Ho Chi Minh City

University of Science

Faculty of Information Technology

Assignment 3
Text-based Search Engine

Introduction to Information Retrieval - CS419

21125148 - Mai Tan Dat

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHTN

ĐHQG-HCM

Ngày 20 tháng 11 năm 2024

1 Functionality

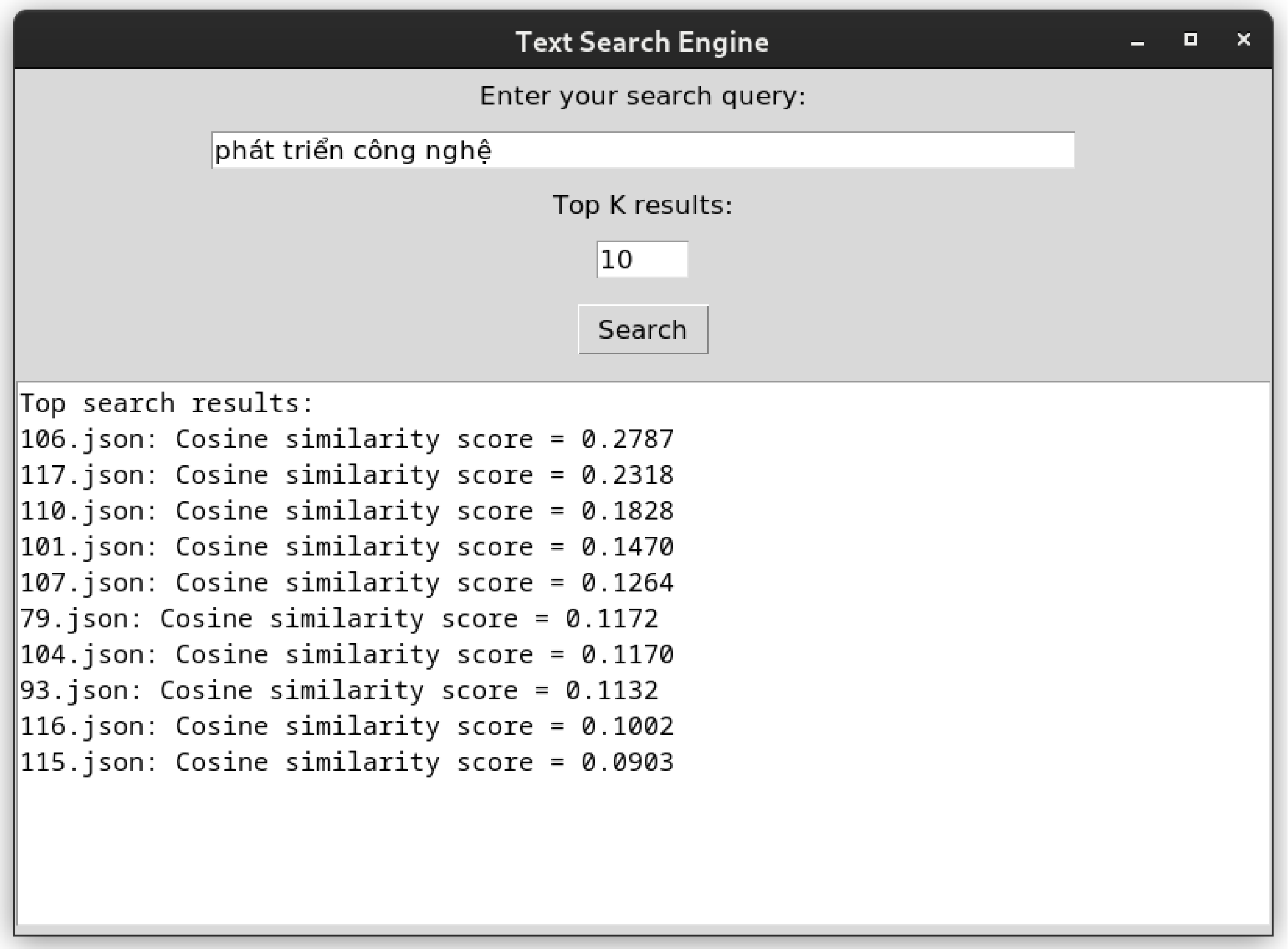
1. Text preprocessing using Underthesea library. Unlike English, Vietnamese is a monosyllabic language so stemming/lemmatization is unnecessary.
    - Text normalization: "oà"→ "òa", "uý"→ "úy", "đột quị"→ "đột quỵ".
    - Word tokenization: "Chàng trai 9X Quảng Trị khởi nghiệp từ nấm sò"→ "Chàng_trai 9X Quảng_Trị khởi_nghiệp từ nấm sò".
    - Stop words removal.
2. TF-IDF indexing using TfidfVectorizer from scikit-learn library.
3. Store the index in json file.
4. Graphical user interface using tkinter library, allow user to enter query string and search for top K relevant documents.
5. Rank documents by cosine similarity.

2 Usage instruction

Put document files in folder "./data".
A document file is a json file with format:

```
{
    "Title": "This is sentence 1",
    "Detail_sapo": "This is sentence 2",
    "Content": [
        "This is sentence 3",
        "This is sentence 4",
        "This is sentence 5",
        ...
    ],
    ...
}
```

In the first run, the program load the data, create document index and save it to the folder "./vectorizer_data". Later, the program will load the saved index.
In Figure 1, when searching for "phát triển công nghệ", the result with the highest cosine similarity is ./data/106.json. In this document, the word "phát triển"appeared 11 times and the word "công nghệ"appeared 24 times.



Enter your search query:

phát triển công nghệ

Top K results:

10

```
Top search results:
106.json: Cosine similarity score = 0.2787
117.json: Cosine similarity score = 0.2318
110.json: Cosine similarity score = 0.1828
101.json: Cosine similarity score = 0.1470
107.json: Cosine similarity score = 0.1264
79.json: Cosine similarity score = 0.1172
104.json: Cosine similarity score = 0.1170
93.json: Cosine similarity score = 0.1132
116.json: Cosine similarity score = 0.1002
115.json: Cosine similarity score = 0.0903
```

Hình 1: Screenshot of the program