Report

1 Approach

1.1 Crawling

Since the collection of text data from the previous A1 assignment may not be sufficient to test the search engine, I decided to crawl more data. The new collection contains 1000 articles from https://tuoitre.vn/:

- 200 articles from the "Thời sự" category
- 200 articles from the "Thế giới" category
- 200 articles from the "Pháp luật" category
- 200 articles from the "Giáo dục" category
- 200 articles from the "Sức khỏe" category

Also, the new data contains one more field, which is the description of the article.

1.2 Indexing

1.2.1 Preprocessing

The first step in indexing is to preprocess the text data. The preprocessing steps include:

1. Normalizing text
2. Removing emoji
3. Removing links
4. Removing punctuation
5. Tokenizing
6. Making lower case
7. Removing stop words

In details:

1. The first step is to normalize the text data, which standardizes the textual data representation. For example, the text "Đảm baỏ chất lượng phòng thí nghiệm hoá học" will be normalized to "Đảm bảo chất lượng phòng thí nghiệm hóa học". This prevents the search engine from treating the same words differently.

2. The second step is to remove emoji from the text data since emoji does not contribute to the meaning of the text and can be removed. This step can be done using regular expressions.
3. The third step is to remove the links since the links in the text data are not useful for searching. This step can also be done using regular expressions.
4. The fourth step is to remove punctuation from the text data since punctuation also does not contribute to the meaning of the text. This step can be done by iterating through each character in the text and removing it if it is a punctuation character. The punctuation characters are defined in the string module.
5. The fifth step is to tokenize the text data, which splits the text data into meaningful units called tokens. The tokens can be words, phrases, or sentences. In this search engine, the tokens are words. The tokenization can be done using the word_tokenize function from the underthesea library. This library is a famous Vietnamese NLP library that provides many useful functions for text processing.
6. The sixth step is to make the text data lower case. This step is important because the search engine should be case-insensitive. For example, the search engine should treat the words "học" and "Học" as the same word.
7. The final step is to remove stop words from the text data. Stop words are common words that not only do not contribute to the meaning of the text but also can cause noise in the search results. For example, the words "là", "và", and "của" are common stop words in Vietnamese. The stop words can be retrieved from an online source.

1.2.2 Creating Index

The index is a data structure that maps terms to documents. The terms are the tokens obtained from the preprocessing step, and the documents are the articles. The index is created using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm. The TF-IDF algorithm is a popular algorithm for information retrieval that assigns weights to terms based on their frequency in the document and their frequency in the corpus.

To do this, TfidfVectorizer class from the sklearn library is used. This class takes a list of text data as input and returns a TF-IDF matrix. The TF-IDF matrix is a matrix where each row represents a document, and each column represents a term. The value in each cell represents the TF-IDF score of the term in the document.

The index is a dictionary that contains two keys: terms and documents. The terms key contains a list of terms, and the documents key contains a list of documents. Each document is a dictionary that contains the document ID and the TF-IDF scores of the terms in the document.

1.2.3 Saving Index

The index is saved to a JSON file for later use. The JSON file contains the index data in a human-readable format. The index can be loaded from the JSON file when the search engine is initialized.

1.3 Searching

1.3.1 Ranking documents

The input query is preprocessed using the same preprocessing steps as the document text. The query vector is then created by setting the elements corresponding to the query terms to 1 and the rest to 0.

The search engine uses the cosine similarity algorithm to rank the documents based on the query. The cosine similarity algorithm is a popular algorithm for information retrieval that measures the similarity between two vectors. In this case, the two vectors are the query vector and the document vector. The query vector is a binary vector that represents the query, and the document vector is a TF-IDF vector that represents the document.

The cosine similarity between two vectors is calculated as follows:

$$\text{similarity} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \times \sqrt{\sum_{i=1}^{n} d_i^2}}$$

where q_i is the i-th element of the query vector, d_i is the i-th element of the document vector, and n is the number of terms.

The search engine calculates the cosine similarity between the query vector and each document vector and ranks the documents based on the similarity score. The documents with higher similarity scores are ranked higher.

1.3.2 Graphical User Interface

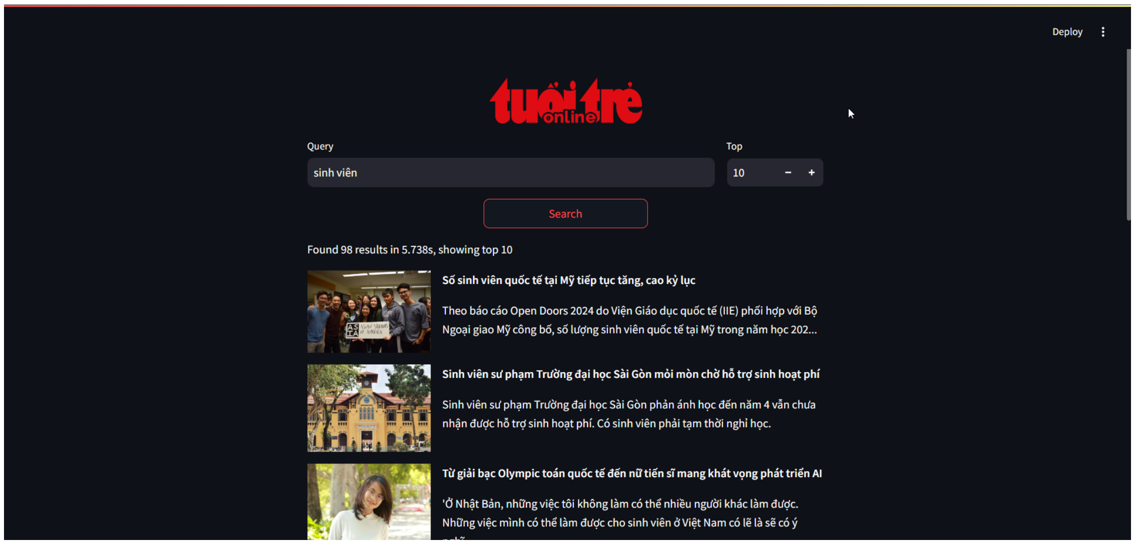
The search engine uses the Streamlit library to create a graphical user interface. The user can enter a query in the input box and click the search button to search for documents. The search results are displayed in a list, with each item containing the thumbnail image, title, and description of the document.

The user can also specify the number of top results to display using a number input box.

2 Challenges

The main challenge I faced is that the computation of cosine similarity is slow for a large number of documents. To speed up the search process, I can use approximate nearest neighbor algorithms such as LSH (Locality Sensitive Hashing) or Annoy (Approximate Nearest Neighbors Oh Yeah). These algorithms can reduce the search time by finding approximate nearest neighbors instead of exact nearest neighbors.

3 Examples



sinh viên

Found 98 results in 5.738s, showing top 10

Số sinh viên quốc tế tại Mỹ tiếp tục tăng, cao kỷ lục

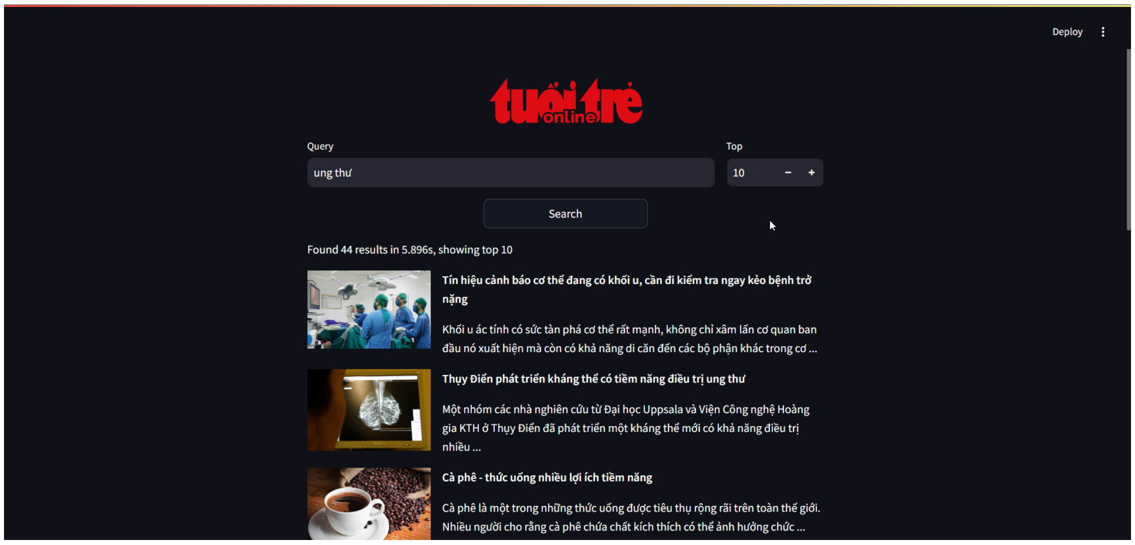
Theo báo cáo Open Doors 2024 do Viện Giáo dục quốc tế (IIE) phối hợp với Bộ Ngoại giao Mỹ công bố, số lượng sinh viên quốc tế tại Mỹ trong năm học 202...

Sinh viên sư phạm Trường đại học Sài Gòn mỏi mòn chờ hỗ trợ sinh hoạt phí

Sinh viên sư phạm Trường đại học Sài Gòn phản ánh học đến năm 4 vẫn chưa nhận được hỗ trợ sinh hoạt phí. Có sinh viên phải tạm thời nghỉ học.

Từ giải bạc Olympic toán quốc tế đến nữ tiến sĩ mang khát vọng phát triển AI

'Ở Nhật Bản, những việc tôi không làm có thể nhiều người khác làm được. Những việc mình có thể làm được cho sinh viên ở Việt Nam có lẽ là sẽ có ý



ung thư

Found 44 results in 5.896s, showing top 10

Tín hiệu cảnh báo cơ thể đang có khối u, cần đi kiểm tra ngay kẻo bệnh trở nặng

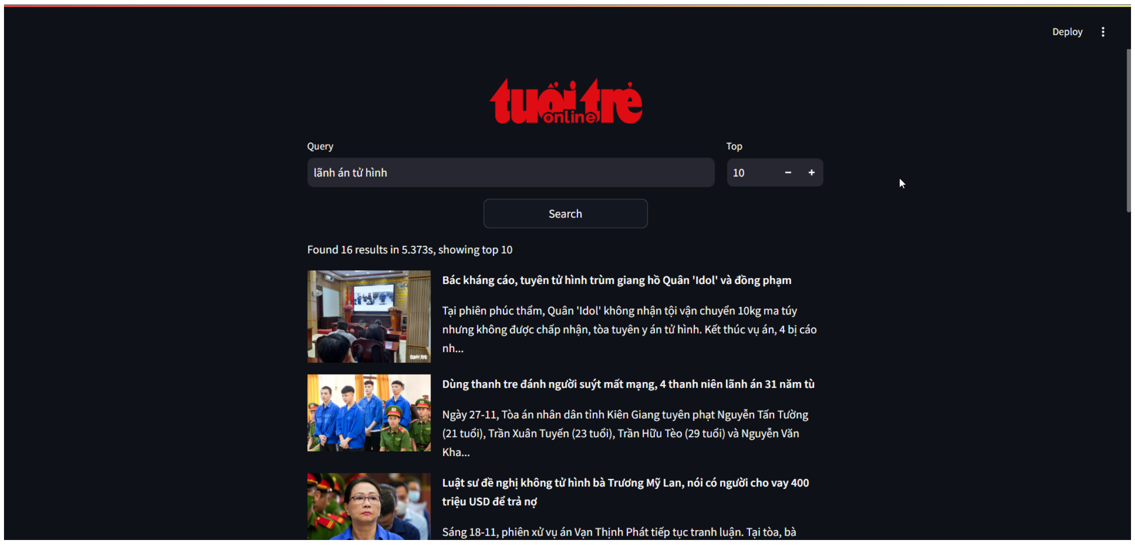
Khối u ác tính có sức tàn phá cơ thể rất mạnh, không chỉ xâm lấn cơ quan ban đầu nó xuất hiện mà còn có khả năng di căn đến các bộ phận khác trong cơ ...

Thụy Điển phát triển kháng thể có tiềm năng điều trị ung thư

Một nhóm các nhà nghiên cứu từ Đại học Uppsala và Viện Công nghệ Hoàng gia KTH ở Thụy Điển đã phát triển một kháng thể mới có khả năng điều trị nhiều ...

Cà phê - thức uống nhiều lợi ích tiềm năng

Cà phê là một trong những thức uống được tiêu thụ rộng rãi trên toàn thế giới. Nhiều người cho rằng cà phê chứa chất kích thích có thể ảnh hưởng chức ...



lãnh án tử hình

Found 16 results in 5.373s, showing top 10

Bác kháng cáo, tuyên tử hình trùm giang hồ Quân 'Idol' và đồng phạm

Tại phiên phúc thẩm, Quân 'Idol' không nhận tội vận chuyển 10kg ma túy nhưng không được chấp nhận, tòa tuyên y án tử hình. Kết thúc vụ án, 4 bị cáo nh...

Dùng thanh tre đánh người suýt mất mạng, 4 thanh niên lãnh án 31 năm tù

Ngày 27-11, Tòa án nhân dân tỉnh Kiên Giang tuyên phạt Nguyễn Tấn Tường (21 tuổi), Trần Xuân Tuyến (23 tuổi), Trần Hữu Tèo (29 tuổi) và Nguyễn Văn Kha...

Luật sư đề nghị không tử hình bà Trương Mỹ Lan, nói có người cho vay 400 triệu USD để trả nợ

Sáng 18-11, phiên xử vụ án Vạn Thịnh Phát tiếp tục tranh luận. Tại tòa, bà

4 References

[1] Vietnamese stopwords: https://github.com/stopwords/vietnamese-stopwords/

[2] Regex for emoji: https://stackoverflow.com/a/58356570/22989557

[3] Regex for url: https://stackoverflow.com/a/17773849/22989557

[4] underthesea library: https://github.com/undertheseanlp/underthesea

[5] TfidfVectorizer class: https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[6] Cosine similarity: https://en.wikipedia.org/wiki/Cosine_similarity