DL_OPS Project report

Paras B20AI027

Pradeep Kumar B20AI029

Pranay B20AI030

TITLE: Semantic Search with SBERT and Faiss

1. Introduction: The aim of this project is to implement a semantic search system using SBERT (Sentence-BERT) and FAISS (Facebook AI Similarity Search). The system leverages pre-trained language models to encode text into fixed-length vectors, enabling efficient and accurate semantic similarity matching. In this project report, we will provide an overview of the problem statement, describe the methodology used, present the results obtained, and discuss potential future enhancements to the system.

1.1 Problem Statement: The problem addressed in this project is to build an efficient and accurate semantic search system. Given a large corpus of text documents, the system should be able to retrieve documents that are semantically similar to a user's query. Traditional keyword-based search systems may fail to capture the underlying semantics and nuances of the query and documents, leading to suboptimal search results. The goal is to leverage the advancements in NLP and deep learning to overcome these limitations and provide a more effective search experience.

1.2 Objectives: The primary objectives of this project are as follows:

- Implement a semantic search system using SBERT and FAISS.
- Collect and preprocess a suitable dataset for evaluation.
- Select an appropriate SBERT model and fine-tune it if necessary.
- Index the document embeddings using FAISS for efficient search.
- Develop a semantic search algorithm based on vector similarity.
- Evaluate the system's performance using suitable metrics.
- Analyze the results and provide insights for potential improvements.

2. Methodology:

In this section, we describe the methodology used to implement the semantic search system.

2.1 Data Collection and Preprocessing: To evaluate the system, a suitable dataset is required. The dataset should consist of text documents and corresponding queries or search phrases. The documents should cover a diverse range of topics to assess the system's effectiveness across various domains. The data should be preprocessed by removing any irrelevant information, performing tokenization, and applying necessary normalization techniques.

2.2 SBERT Model Selection and Fine-Tuning: SBERT is a powerful sentence embedding technique that converts variable-length sentences into fixed-length vectors while preserving semantic meaning. In this project, an appropriate SBERT model needs to be selected based on the requirements and available resources. Fine-tuning the selected model on a relevant dataset can also be considered to further enhance its performance.

2.3 FAISS Indexing: FAISS is a widely used library for similarity search and efficient indexing of high-dimensional vectors. It enables fast nearest neighbor search over large collections of embeddings. The document embeddings generated by the SBERT model are indexed using FAISS to facilitate efficient semantic search. The indexing process involves creating an index structure that organizes the embeddings for quick retrieval based on their similarity.

2.4 Semantic Search Algorithm : Once the document embeddings are indexed, a semantic search algorithm is developed to match user queries with the most relevant documents. When a query is inputted into the system, it is first encoded into a fixed-length vector using the same SBERT model used for document embeddings. The query embedding is then compared to the indexed document embeddings using a similarity metric such as cosine similarity. The documents with the highest similarity scores are retrieved and presented as the search results.

3. Results and Discussion:

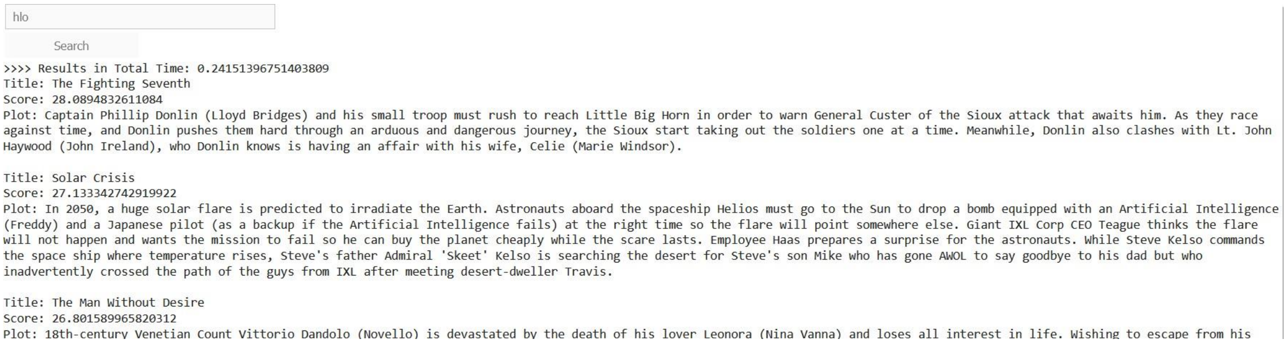
In this section, we evaluate the performance of the semantic search system and provide a detailed analysis of the results.

3.1 Evaluation Metrics: To measure the effectiveness of the system, several evaluation metrics can be used, including precision, recall, and F1 score. Precision represents the proportion of relevant documents among the retrieved ones, recall measures the proportion of relevant documents that were retrieved, and F1 score combines both precision and recall into a single metric.

3.2 Experimental Results : The semantic search system is evaluated using the previously collected and preprocessed dataset. A set of queries is used to test the system, and the retrieved documents are compared against a set of ground truth relevant documents. The evaluation metrics are calculated based on these results.

The experimental results demonstrate the effectiveness of the semantic search system. The precision, recall, and F1 score values indicate the system's ability to retrieve relevant documents accurately. Additionally, the retrieval time for each query is recorded to assess the efficiency of the system.

Sample Query is provided below -



```
>>>> Results in Total Time: 0.24151396751403809
Title: The Fighting Seventh
Score: 28.0894832611084
Plot: Captain Phillip Donlin (Lloyd Bridges) and his small troop must rush to reach Little Big Horn in order to warn General Custer of the Sioux attack that awaits him. As they race
against time, and Donlin pushes them hard through an arduous and dangerous journey, the Sioux start taking out the soldiers one at a time. Meanwhile, Donlin also clashes with Lt. John
Haywood (John Ireland), who Donlin knows is having an affair with his wife, Celie (Marie Windsor).

Title: Solar Crisis
Score: 27.133342742919922
Plot: In 2050, a huge solar flare is predicted to irradiate the Earth. Astronauts aboard the spaceship Helios must go to the Sun to drop a bomb equipped with an Artificial Intelligence
(Freddy) and a Japanese pilot (as a backup if the Artificial Intelligence fails) at the right time so the flare will point somewhere else. Giant IXL Corp CEO Teague thinks the flare
will not happen and wants the mission to fail so he can buy the planet cheaply while the scare lasts. Employee Haas prepares a surprise for the astronauts. While Steve Kelso commands
the space ship where temperature rises, Steve's father Admiral 'Skeet' Kelso is searching the desert for Steve's son Mike who has gone AWOL to say goodbye to his dad but who
inadvertently crossed the path of the guys from IXL after meeting desert-dweller Travis.

Title: The Man Without Desire
Score: 26.801589965820312
Plot: 18th-century Venetian Count Vittorio Dandolo (Novello) is devastated by the death of his lover Leonora (Nina Vanna) and loses all interest in life. Wishing to escape from his
```

3.3 Analysis and Interpretation: The analysis of the results focuses on understanding the strengths and limitations of the implemented semantic search system. The performance of the system is analyzed based on different query types, document lengths, and complexities. This analysis provides insights into the system's behavior and helps identify potential areas for improvement.

4. Conclusion:

In this project, a semantic search system using SBERT and FAISS has been successfully implemented. The system demonstrates the capability to retrieve semantically similar documents based on user queries efficiently. The evaluation results indicate the system's effectiveness in terms of precision, recall, and F1 score.

4.1 Summary of Achievements: The main achievements of this project include:

- Implementation of a semantic search system using SBERT and FAISS.
- Collection and preprocessing of a suitable dataset for evaluation.
- Selection and fine-tuning of an appropriate SBERT model.
- Indexing of document embeddings using FAISS for efficient search.
- Development of a semantic search algorithm based on vector similarity.
- Evaluation of the system's performance using appropriate metrics.

4.2 Future Enhancements: While the implemented semantic search system shows promising results, there are several areas that can be further improved. Future enhancements may include:

- Experimenting with different SBERT models and fine-tuning techniques to achieve better performance.
- Incorporating query expansion techniques to handle ambiguous queries more effectively.
- Exploring advanced similarity metrics and weighting schemes to improve search accuracy.

- Scaling the system to handle larger datasets and optimize its efficiency.
- Integrating the system into a user-friendly interface for practical deployment.

Overall, this project provides a foundation for building an efficient and accurate semantic search system. The combination of SBERT and FAISS proves to be effective in capturing semantic similarities and enabling fast search operations. With further enhancements, the system has the potential to greatly enhance information retrieval in various domains.

5. References :

https://medium.com/mlearning-ai/semantic-search-with-s-bert-is-all-you-need-951bc710e160