CY SHIELD

Semantic Search in articles using NLP

Overview

The goal of this project is to implement a document similarity search system using FAISS (Facebook AI Similarity Search) and Sentence-BERT (Sentence Embeddings using BERT). The system takes a user query and retrieves the most similar article from the **abc_news** dataset based on their semantic similarity.

About Data Set

This contains data of news headlines published over a period of nineteen years.Sourced from the reputable Australian news source ABC (Australian Broadcasting Corporation)

Agency Site: (http://www.abc.net.au)

It includes the entire corpus of articles published by the abcnews website in the given date range.

Content

Format: CSV ; Single File

1. **publish_date:** Date of publishing for the article in yyyyMMdd format
2. **headline_text:** Text of the headline in Ascii , English

Data contains 85041 entries ,each of which is a news

Baseline Experiments

The objective of this project is to develop a search engine that combines FAISS (Facebook AI Similarity Search) and Sentence-BERT (Sentence Embeddings using BERT). Which aims to search some of words in English articles then extract hot keywords and similar articles to the query the user enters

1. Data Preprocessing:

- The "abcnews-date-text" dataset, which comprises news headlines and their corresponding publication dates, is loaded.

- Text preprocessing techniques are applied to ensure data cleanliness and standardization.

- These techniques involve:

  Removing punctuation: eliminating punctuation marks from the text data before further analysis or processing. Punctuation marks include characters such as commas, periods, exclamation marks, quotation marks. regular expressions (regex) are used to match specific characters or character groups in a string. In this case, the regular expression pattern is used to identify any characters that are not alphanumeric (letters or numbers) or whitespace

  Converting text to lowercase: By converting all characters to lowercase, we ensure that the same word or term, regardless of its capitalization, is treated as identical throughout the text .Also we ensure compatibility with these pre-trained models and facilitate better word representation learning.

  Eliminating stop words:The goal of eliminating stop words is to remove these common words from the text before further analysis. By doing so, we can focus on the more meaningful and content-bearing words in the text,
  performing word stemming using the Porter stemmer algorithm.Examples of stopwords in English include words like "the," "and," "is," "are," "in," "to," etc.

Baseline Experiments (continue)

Stemming: The stem represents the core meaning of a word and is obtained by removing prefixes, suffixes, and other word endings. The goal of stemming is to group together words that have the same root, despite their different inflections or derivations.

2. Sentence-BERT Encoding

Sentence-BERT (Bidirectional Encoder Representations from Transformers) is a technique used for encoding and representing sentences or short texts in a dense vector space. It is based on the Encoder Transformer architecture.

SBERT uses the encoder part in the transformers to capture the semantic meaning and contextual information of sentences by mapping them to fixed-length vector representations. Unlike traditional word-level embeddings, Sentence-BERT takes into account the entire sentence and considers the context and surrounding words when generating the embeddings and it is widely used in semantic similarity.

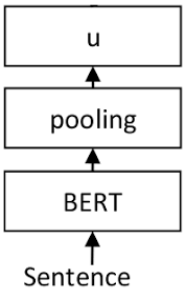
Model Architecture :

1. **Pretrained Transformer Mode**l:It uses BERT to capture contextual information and semantic relationships between words.

2. **Tokenization**: The input sentences are tokenized, breaking them down into individual words (tokens) . Each token is assigned a unique ID.

3. **Word Embeddings**:The tokens are then transformed into word embedding (vectors),which captures the contextual information and meaning of each word based on its surrounding words in the sentence.

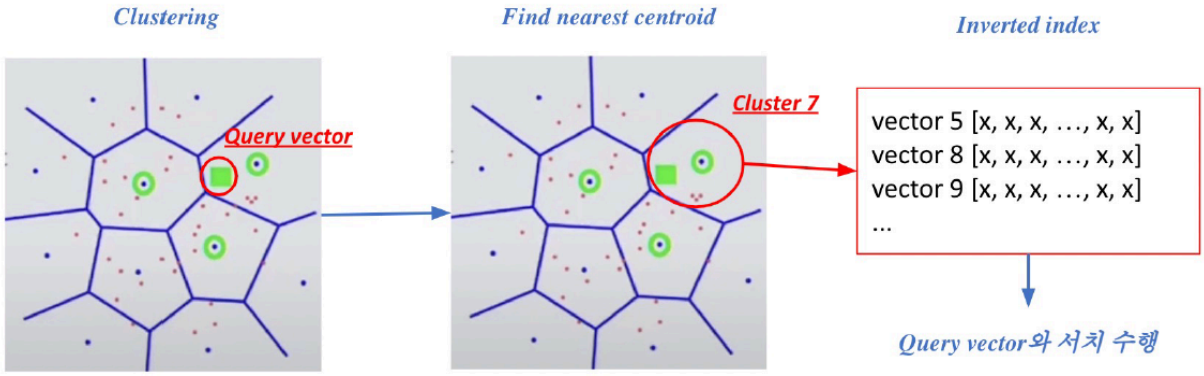
4. **Sentence Embedding:** SBERT applies Pooling for summarizing the information captured by the individual word embeddings into a compact and fixed-length representation of the entire sentence



3.FAISS(Facebook AI Similarity Search) :

FAISS (Facebook AI Similarity Search) is a library developed by Facebook AI Research that specializes in efficient similarity search and clustering of large-scale datasets. It is widely used for handling high-dimensional vector spaces. It offers two primary indexing methods: IndexFlat and IndexIVF. IndexFlat is a straightforward approach that searches the entire dataset, while IndexIVF (Inverted File) partitions the data into smaller cells or **clusters** for faster search.

IndexIVF (Inverted File) : The IndexIVF method involves partitioning the **abc_news** dataset into smaller cells or clusters in the **vector space** to speed up the search process. Each cell is represented by an inverted file, which is essentially an index structure that maps each vector to a specific cell. The inverted file stores a list of vectors associated with each cell.



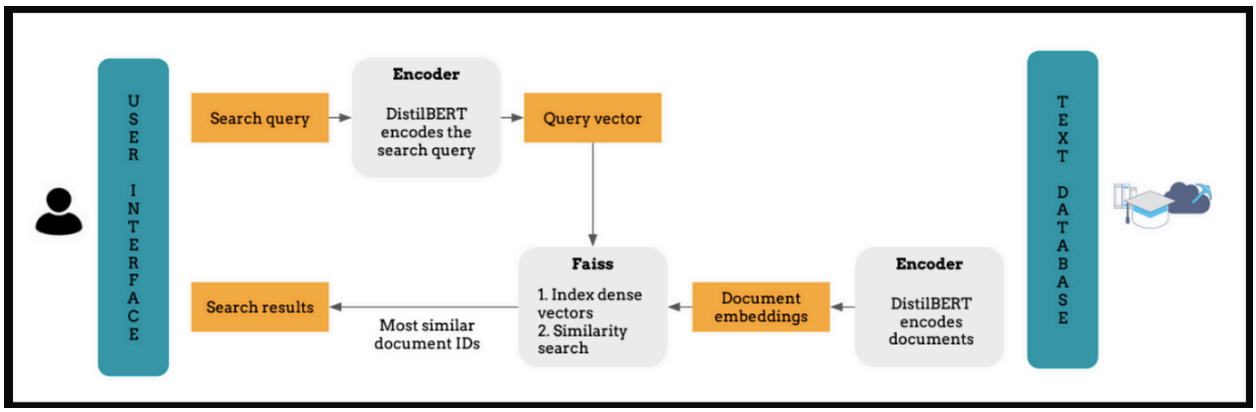
4.FAISS+SBERT :

Here's how FAISS + SBERT works together:

1. **Sentence-BERT Encoding**: First, the text data undergoes preprocessing steps explained above.This preprocessed text is then encoded using the Sentence-BERT model, which generates dense vector representations, known as sentence embeddings, for each input sentence.

2. **FAISS Indexing**: Once the sentence embeddings are obtained, FAISS comes into play. FAISS provides efficient indexing with IndexIVFFlat, which creates an index structure that organizes the sentence embeddings, making it easier and faster to perform similarity search.

3. **Indexing and Searching**: The sentence embeddings are added to the FAISS index, associating each embedding with a unique identifier or index. The index structure is then trained and constructed based on the embeddings, optimizing it for similarity search. Once the index is ready, user queries can be entered, which undergo the same preprocessing and encoding steps as the original data. The query embeddings are then compared against the indexed embeddings using FAISS's efficient search algorithms, retrieving the **most similar articles and hot words based** on their semantic similarity.



End results:

Here the entered query in “School”
Viewing only top 5 similar articles
Total time it took the engine to retrieve relevant data in milliseconds
The similarity measure here is from the distance matrix between vectors
Viewing hot words

```
Enter your query: school
Total time: 0.0314798355102539 06
Results:
Result: sars threat closes beijing schools
Similarity Score: -221.7443
Publication Date: 20030423
Hot Word: school

Result: forums to consider school leaving age
Similarity Score: -219.3739
Publication Date: 20040209
Hot Word: school

Result: education minister raises school leaving age at
Similarity Score: -214.3020
Publication Date: 20040211
Hot Word: school

Result: pm encourages debate over schools
Similarity Score: -213.1128
Publication Date: 20040126
Hot Word: school

Result: union highlights school concerns
Similarity Score: -212.8883
Publication Date: 20030808
Hot Word: school
```

Here the entered query in "How high school football coach is focusing on mental health"
Viewing only top 5 similar articles
Total time it took the engine to retrieve relevant data in milliseconds
The similarity measure here is from the distance matrix between vectors

Viewing hot words

```
Enter your query: How high school football coach is focusing on mental health
Total time: 0.028017282485961914
Results:
Result: drivers urged to take care as school resumes
Similarity Score: -164.4830
Publication Date: 20031006
Hot Word: school

Result: detainee children will attend port augusta schools
Similarity Score: -162.4056
Publication Date: 20030226
Hot Word: school

Result: parents teachers await schools review
Similarity Score: -161.0366
Publication Date: 20040415
Hot Word: school

Result: child detainees to attend primary school in sa
Similarity Score: -159.8206
Publication Date: 20030306
Hot Word: school

Result: support for driver ed school to continue
Similarity Score: -158.2196
Publication Date: 20030605
Hot Word: school
```

Here the entered query in "Kourtney Kardashian announces she is pregnant with Travis Barker at Blink-182 concert"
Viewing only top 5 similar articles
Total time it took the engine to retrieve relevant data in milliseconds
The similarity measure here is from the distance matrix between vectors
Viewing hot words

```
Enter your query: Kourtney Kardashian announces she is pregnant with Travis Barker at Blink-182 concert
Total time: 0.03139042854309082
Results:
Result: scream queen sharapova shrieks into semi finals
Similarity Score: -202.9151
Publication Date: 20030614
Hot Word: queen

Result: pop star anastacia recalls breast cancer shock
Similarity Score: -200.5634
Publication Date: 20040331
Hot Word: queen

Result: britney spears seeks annulment of vegas wedding
Similarity Score: -192.7312
Publication Date: 20040106
Hot Word: queen

Result: agassi advances serena smashes schett
Similarity Score: -189.6812
Publication Date: 20030531
Hot Word: queen

Result: queen leaves hospital after double surgery
Similarity Score: -189.3689
Publication Date: 20031213
Hot Word: queen
```

Other Experiments

Another approach to calculate similarity between user query and articles is by **Cosine Similarity. The output was more understandable for user because similarity values ranges from 0 to 1 .However this cosine similarity took more time to find similar articles than distance matrix.**

Here the entered query in "How high school football coach is focusing on mental health"
Viewing only top 5 similar articles
Total time it took the engine to retrieve relevant data in : 35.05 seconds
The similarity measure here is from the distance matrix between vectors
Viewing hot words

```
Enter your query: How high school football coach is focusing on mental health
Total time: 35.056047201156616
Results:
Similarity Score: 0.2944
Publication Date: 20031006
drivers urged to take care as school resumes

Similarity Score: 0.3653
Publication Date: 20030226
detainee children will attend port augusta schools

Similarity Score: 0.3878
Publication Date: 20040415
parents teachers await schools review

Similarity Score: 0.3174
Publication Date: 20030306
child detainees to attend primary school in sa

Similarity Score: 0.2911
Publication Date: 20030605
support for driver ed school to continue
```

Overall Conclusion:

The document similarity search system implemented using FAISS and Sentence-BERT provides a solution for retrieving similar articles based on user queries. The system leverages preprocessed and encoded data along with efficient indexing to achieve fast and accurate search results. Further experiments can be conducted to enhance the system's performance and explore different techniques for measuring similarityl.

Tools Used:

- Python programming language
- FAISS for similarity search
- SentenceTransformer for Sentence-BERT
- Pandas for data manipulation
- NumPy for numerical operations
- NLTK for text preprocessing
- Scikit-learn for CountVectorizer
- Colabr Notebook for code execution and documentation

Challenges Faced:

The biggest challenge faced in this project may be the optimization of the search process. Finding the right balance between indexing techniques, preprocessing methods, and the choice of the Sentence-BERT model can be challenging to achieve optimal performance.

Key Learnings:

How to leverage 2 model with each other ,which are the sbert and faiss ,to perform a specific task