Machine Learning Project Presentation

Ishan Upadhyaya (112115064)
Varshil Kavathiya (112115071)
Arvind Khandelwal (112115073)
Mansi Singh (112115086)
Department of Computer Science

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY PUNE
Leading Technology.
Leveraging Technology.

HAVE YOU EVER READ A BOOK AND FOUND THAT THIS BOOK WAS SIMILAR TO ANOTHER BOOK THAT YOU HAD READ BEFORE? I HAVE ALREADY. PRACTICALLY ALL SELF-HELP BOOKS THAT I READ ARE SIMILAR TO ENID BLYTON’S BOOK.

SO WE WONDERED IF NATURAL LANGUAGE PROCESSING COULD MIMIC THIS HUMAN ABILITY AND FIND THE SIMILARITY BETWEEN DOCUMENTS.

Document Similarity using Natural Language Processing

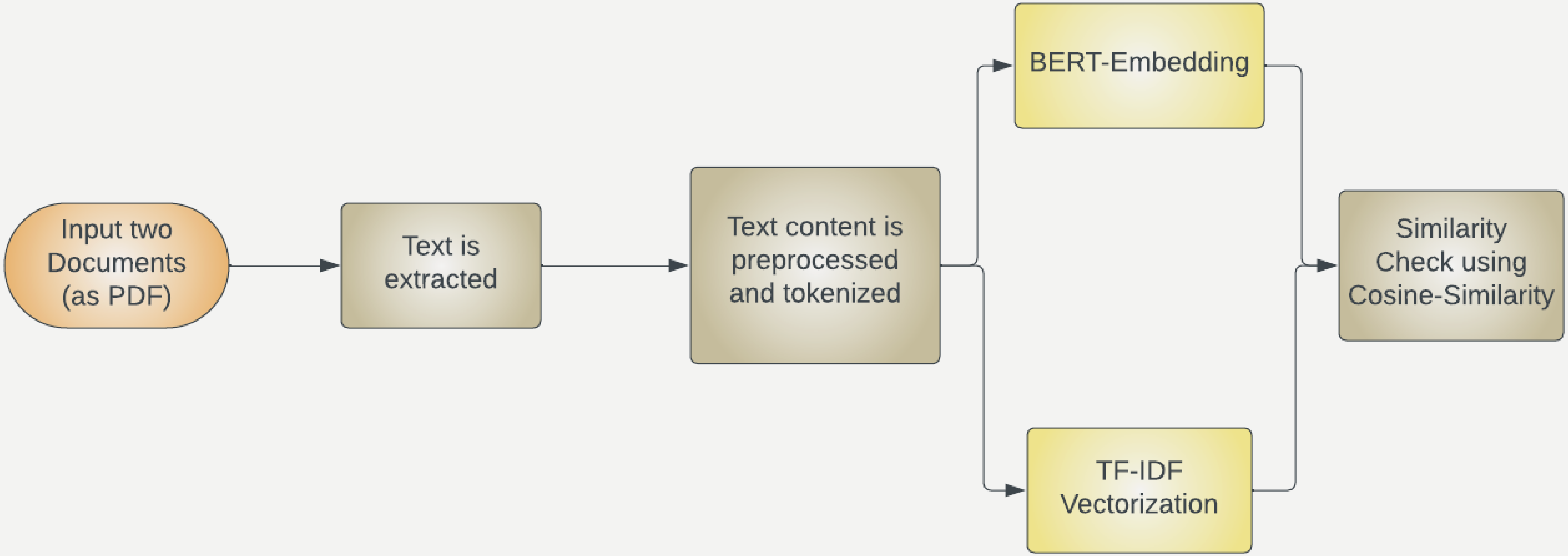
OVERVIEW OF WORK

The Document Similarity project leverages Natural Language Processing (NLP) techniques to compare the similarity between two documents in PDF format.

The goal of this project is to provide a tool for users to determine how closely related or similar two documents are by analyzing their textual content.

It utilizes two main approaches: **TF-IDF** and **BERT**-based embeddings, combined with cosine similarity.

FLOWCHART



Input two Documents (as PDF)

Text is extracted

Text content is preprocessed and tokenized

BERT-Embedding

TF-IDF Vectorization

Similarity Check using Cosine-Similarity

OVERVIEW OF TF-IDF(TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY):

This approach is based on the frequency of terms within the documents and their significance across the entire collection of documents.

TF-IDF calculates a numeric score for each term in a document, reflecting its importance.

The cosine similarity is then used to compare the TF-IDF vector representations of two documents, determining their similarity.

Term Frequency: TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Inverse Document Frequency: IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

The TF-IDF of a term is calculated by multiplying TF and IDF scores.

$$TF\text{-}IDF = TF * IDF$$

OVERVIEW OF BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS):

BERT is a deep learning model that considers the context and meaning of words within documents.

BERT embeddings are used to represent the documents as dense vectors, capturing the semantic relationships between words.

The cosine similarity is applied to compare these BERT embeddings, providing a context-aware measure of document similarity.

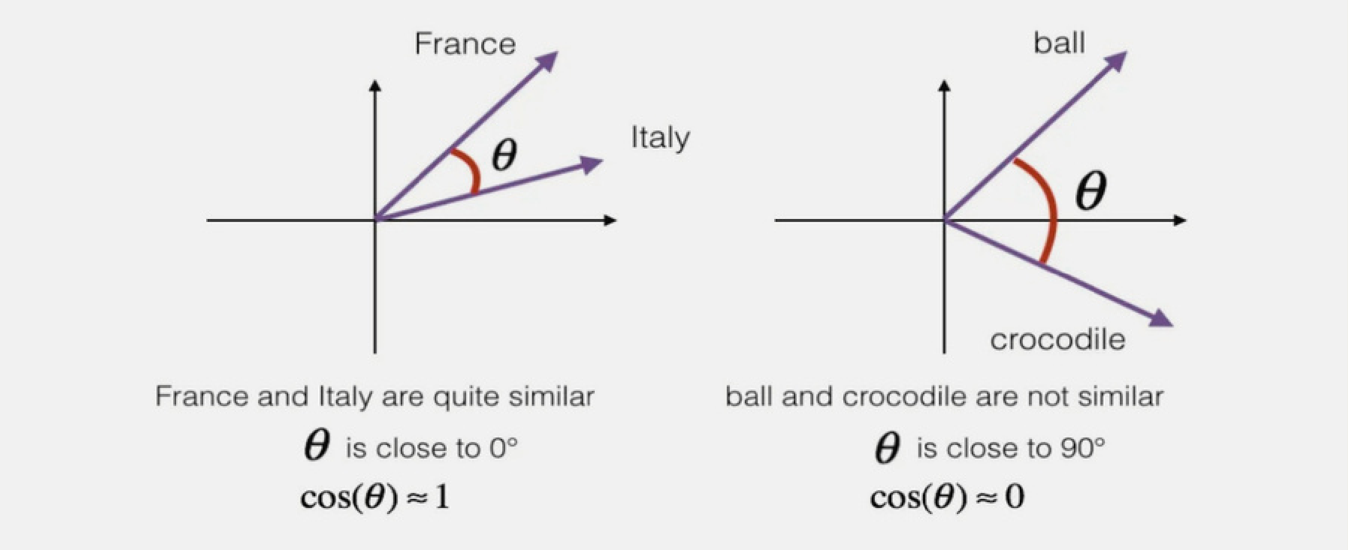
USING COSINE SIMILARITY

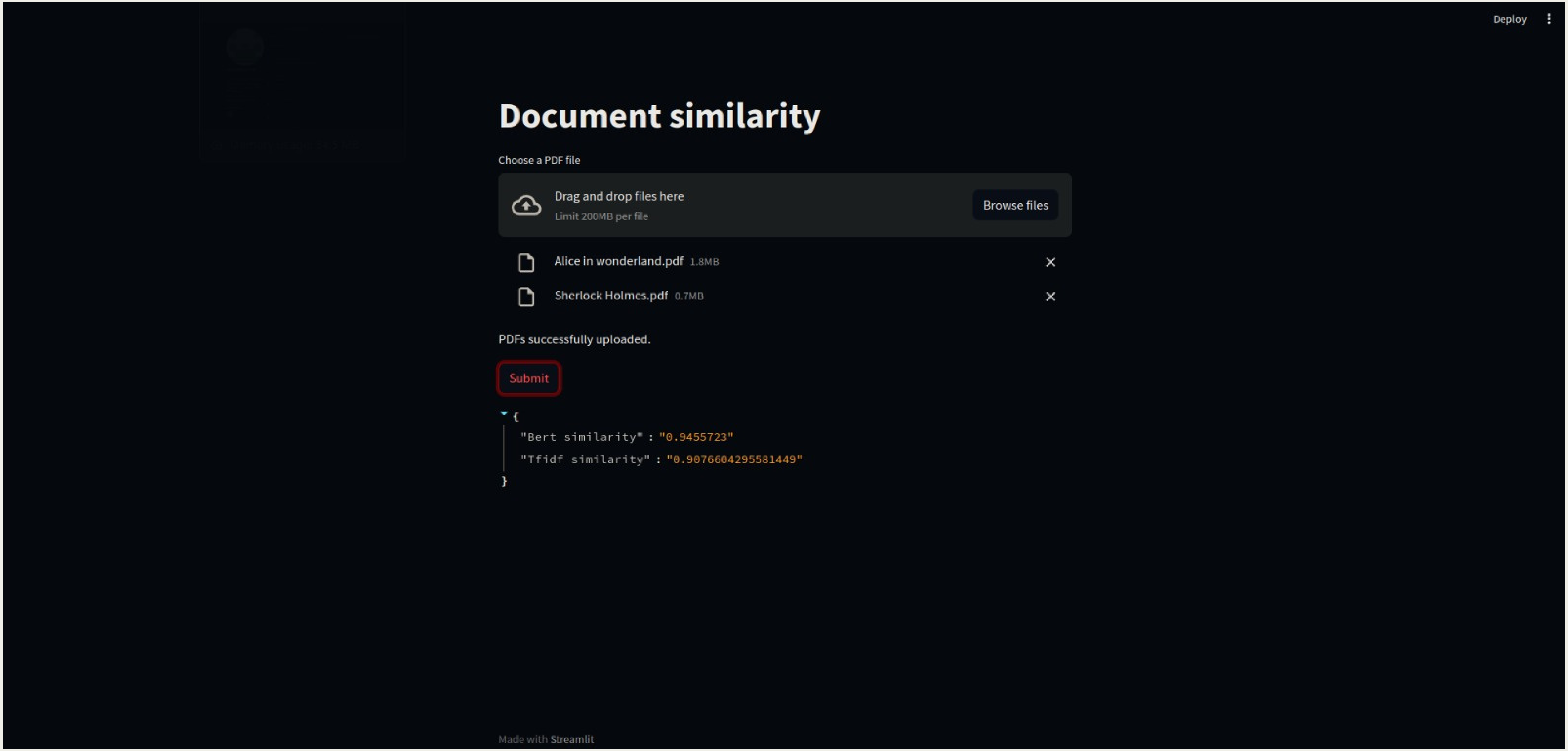
Cosine similarity is a mathematical measure used to quantify the similarity between two vectors.

In the context of document similarity, it is widely used to compare document vectors created by TF-IDF or BERT.

It compares the TF-IDF vectors or BERT embeddings of two documents and provides a numerical score representing their similarity.

A higher cosine similarity score indicates greater document similarity, while a lower score indicates dissimilarity.





Document similarity

Choose a PDF file

Drag and drop files here

Limit 200MB per file

Alice in wonderland.pdf 1.8MB

Sherlock Holmes.pdf 0.7MB

PDFs successfully uploaded.

```
{
  "Bert similarity" : "0.9455723"
  "Tfidf similarity" : "0.9076604295581449"
}
```

Made with Streamlit

Thank you!