CS59001 DESIGN LABORATORY PROPOSAL

Neural File Search Engine

Abhinav Bohra, 18CS30049

Introduction

Semantic search [1] is a data-searching technique that goes beyond simply looking for keywords to understand the user's intent and the context of the words they are using.

Windows Search [2] is integrated into all Windows Explorer windows for instant access to search. This enables users to quickly search for files and items by file name, properties, and full-text contents. The search technique is based on fuzzy search and text matching with different file properties. However, it does not consider the language semantics of file's name and its content.

At its core, Search is about understanding language. With introduction of Bidirectional Encoder Representations from Transformers (BERT) [3], a model that process words in relation to all the other words in a sentence, rather than one-by-one in order, it is now possible to consider the full context of a word by looking at the words that come before and after it which is particularly useful for understanding the intent behind search queries.

Problem Statement

The objective is to design and develop an NLP based intelligent local-file search engine that searches for relevant text documents in a specific folder, considering the semantics of the file's name & its content and returns the most relevant files.

Input

- **Search Directory:** Location of folder to be searched
- **Search Query:** Phrase/keywords to be searched

Output

- **Relevant Files:** Location of top matched files

Files Supported

- **File Content:** .docx, .txt
- **File Name:** .docx, .txt, .pdf, .csv, .xlsx, jpg, .png, .mp3, .mp4

Proposed Design

As shown in Figure 1, the tool takes directory path and query phrase as input and outputs the relevant documents.

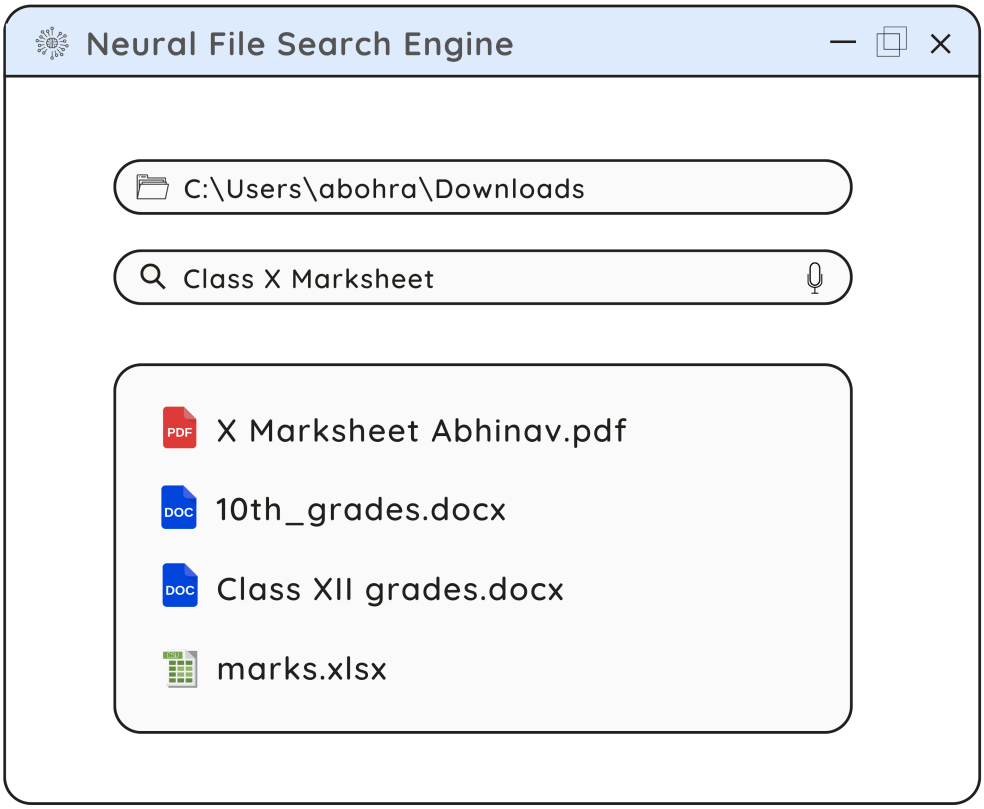

Figure 1: User Interface Mock-up

Solution Approach

The problem can be divided into following subproblems:-

1. Read Input: Read and parse documents to extract the text content.

2. Text Processing: Perform basic text preprocessing steps such as cleaning, stemming, spelling correction on both file content as well as the search query phrase. Split documents into equisized chunks and use average feature representation, to process longer documents.

3. Create Embeddings: Obtain sentence level and word level latent representation of documents and query phrase respectively using a pre-trained language model such as BERT [3].

4. Matching: Find documents similar to query using cosine similarity and display file path of document(s) having score greater than a set probability threshold.

Future Work

In future, we can extend this tool to a multi-modal search engine that supports image, audio and video data by utilising Vision Transformers and Automatic Speech Recognition. By using pre-trained language models like mBERT and XLM [4], we can enable multilingual search queries as well.

References

[1] T. Roberts, "Bloomreach," 2019, https://www.bloomreach.com/en/blog/2019/semantic-search-explained-in-5-minutes.

[2] Microsoft, 2022, https://docs.microsoft.com/en-us/windows/win32/search/-search-3x-wds-overview.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.