TextSeek

Aadil Hoda 160101001
Shaurya Gomber 160101086
Rishabh Jain 160101088

Introduction

TextSeek is an easy-to-use file search python program that includes advanced functionality for power users! TextSeek uses Vector Space Model to rank-order the documents matching a query.

Basic Features

SEARCH IN FILE

- User can search for a query in all the files in a directory.

RANKING

- The search results are ranked according to their match percent.
- Top 10 results are displayed at a time, with an option to view next 10.

Extra Features

FOLDER INPUT

- User inputs the directory name where the files are to be searched.

THREE SEARCH MODES

- Any : Displays files containing any word in query.
- All : Displays only files containing all the words in the query .
- Exact : Displays only files containing the words in query in exact order.

FILE OPENING

- User can open a file in the search results by inputting the serial number.
- On opening a file, line numbers are printed and the words in the query are highlighted.
- User can edit the file and the search results will be updated on next query.

Data Structures and Algorithms

DATA STRUCTURES USED

- Lists
- Dictionaries

ALGORITHMS

- Vector Space Model
  Documents and queries are represented as vectors. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Dot product of the unit vectors of query and file vectors are used for match score calculation.
- Henry Spencer Regex Library
  We are using Vim text editor which uses this library for pattern matching. The implementation is through regular expressions using nondeterministic automata which makes it simple, small, and fast at compiling regular expressions.

Time Complexity

Python Lists – Append, Get Value, and Get Length	O(1)
Python Dictionary – Insert and Get Item *(using Hashing)*	O(1)

PREPROCESSING

We create a list of dictionaries, such that dictionary of each file contains the unique words in the file and their frequencies. O(C)
where C is the total number of characters in all the files.

QUERY (ANY OR ALL SEARCH)

We make a dictionary of the query. If the query length is L, O(L)
For each file in the folder, we take the dot product of the query and the file's dictionary vectors. O(NL)
where N is the number of files in the folder

QUERY (EXACT SEARCH AND HIGHLIGHTING)

On the file set A returned from the all search, we use the Vim text editor which implements the Henry Spencer Regex Library. $O(NL+C_A)$
where C_A is the number of characters in file set A.