COL106 : Data Structures and Algorithms

Assignment 7 : Corpus Q&A Tool

Team – stepbrothers

Soumya Namdeo, 2022MT61978

Lokendra Singh Gohil, 2022EE11164

Deevyansh Khadria, 2022EE31883

Arpit Mourya, 2022EE11728

Here, we have stored data using Hashtables containing two classes: "data" and "element". Data represents a unique paragraph, and the element class stores a word and a vector that stores its location using data type pointers. The size of this vector in the element class gives us the frequency of a word in the whole document. We used hashing methods on these 2 data types to store data efficiently. We then proceeded to calculate the scores of words and paragraphs. We then took the paragraphs for a given input (after removing stop words which are more frequently used from the input statement) statement and sorted them on the basis of their page and book indexing because it is more likely to find answers from neighbouring paragraphs. We are also prioritizing the paragraphs having more than one keyword from the input query.

From here, we moved on to hubbing. In hubbing, we grouped the sorted data pointers (on the basis of page and book code indexing) obtained from a given query input on the basis of criteria according to which the page difference between page numbers of two successive data pointers was maintained to be less than some constant.

For example – if we have

Unsorted search results related to the query :- (1,1,1), (3,1,1), (2,1,2), (1,1,2), (2,2,2), (5,1,4), (5,3,4).

Sorted search results :- (1,1,1), (1,1,2), (2,1,2), (2,2,2), (5,1,4), (5,3,4).

After hubbing :- Hub1 :- (1,1,1), (1,1,2)

Hub2 :- (2,1,1), (2,2,2)

Hub3 :- (5,1,4), (5,3,4)

We calculated the hub score of every hub which is the summation of scores of all paragraphs in that hub and sorted these hubs on the basis of their hub scores. By doing this, we got results that are close to each other such that more relevant paragraphs appear on the top together. Then we passed the top few paragraphs from the top hubs to ChatGPT AI and got the response.