EvaDB Assignment 1 Report: AI Search Engine

Anmol Agarwal (aagarwal622)

Link: https://github.com/anmolagarwalcp810/CS6422-EvaDB-AI_Search_Engine

Note: Earlier code was hosted on https://github.gatech.edu/aagarwal622/AI_Search_Engine, and has been moved to https://github.com/anmolagarwalcp810/CS6422-EvaDB-AI_Search_Engine.

Overview:
In this project, I have implemented a search engine using AI. Here the user will enter a query, and will get the relevant paragraphs from documents in the output based on the similarity score, i.e., those paragraphs which have the highest similarity scores would be returned the user. User also has the ability to get the summary of responses to their query. This summary would generated by Facebook-bart-cnn language model. Furthermore, they can also control how many relevant paragraphs does the user want to see in the output. Moreover, the search engine will also regularly update itself as new documents are added or deleted from the directory.

Versions:
Python: 3.11.4
EvaDB: 0.3.7
OS: Ubuntu 22.04.3 LTS

Features:
The AI Search Engine has the following features:
1. Return relevant paragraphs to the user most similar to the query. Currently we support both PDF and text documents.
2. Also give a summary response to the user (if they need it) by consolidating all the relevant paragraphs and generating a summary out of them. User can toggle this feature using "ENABLE/DISABLE SUMMARY" commands.
3. By default, at most 10 paragraphs are returned to the user, but they can change this limit by running LIMIT <new_limit> command to control how many results do they want.
4. If user wants to see the backend table storing all the documents, they can run SHOW command.
5. Finally, when users add or delete some documents from the directory to which the search engine database is pointing to, it will automatically update itself accordingly, i.e., add new rows for new documents or delete rows from itself for deleted documents. The engine will check the directory roughly once every 60 seconds or so.

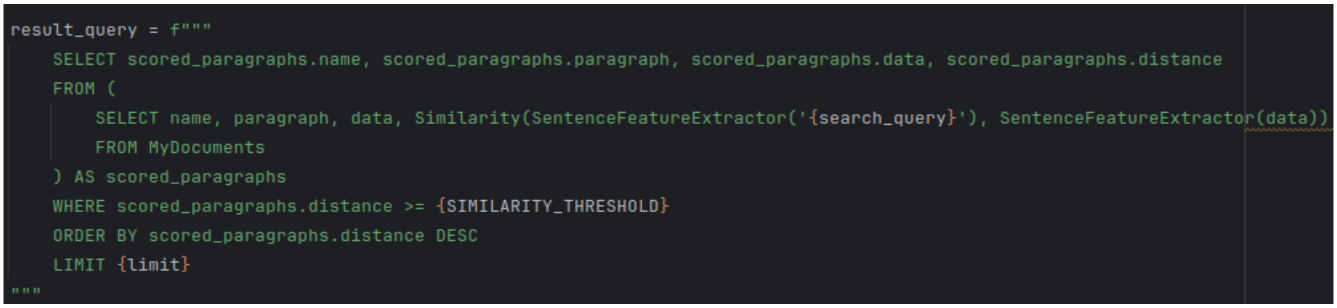
Please refer demo section to see all the above-mentioned features at play.

Important Files and Folders:
Main program: ai_search_engine.py
Main folder: docs (contains all the documents to used by our search engine)

How to Run:
> python ai_search_engine.py

Implementation Details:

1. First I loaded PDFs and Text files from directory. PDFs were loaded into a separate table using “LOAD PDF” command and text files were read and divided into paragraphs split by newline character and inserted into another table.
2. The table for PDFs is called MyPDFs and table for text files is called MyDocuments.
3. To simplify further process, I copied all the entries from MyPDFs into MyDocuments table and removed MyPDFs table. MyDocuments contains all the documents with each row corresponding to a paragraph.
4. Then I created sentence feature extractor from abstract function implementation stored in functions/sentence_feature_extraction.py (this was taken from EvaDB github repository). I also created vector index.
5. Now the user enters the query, and I use the following query to first get embeddings, and get similarity between query and entries in MyDocuments table.



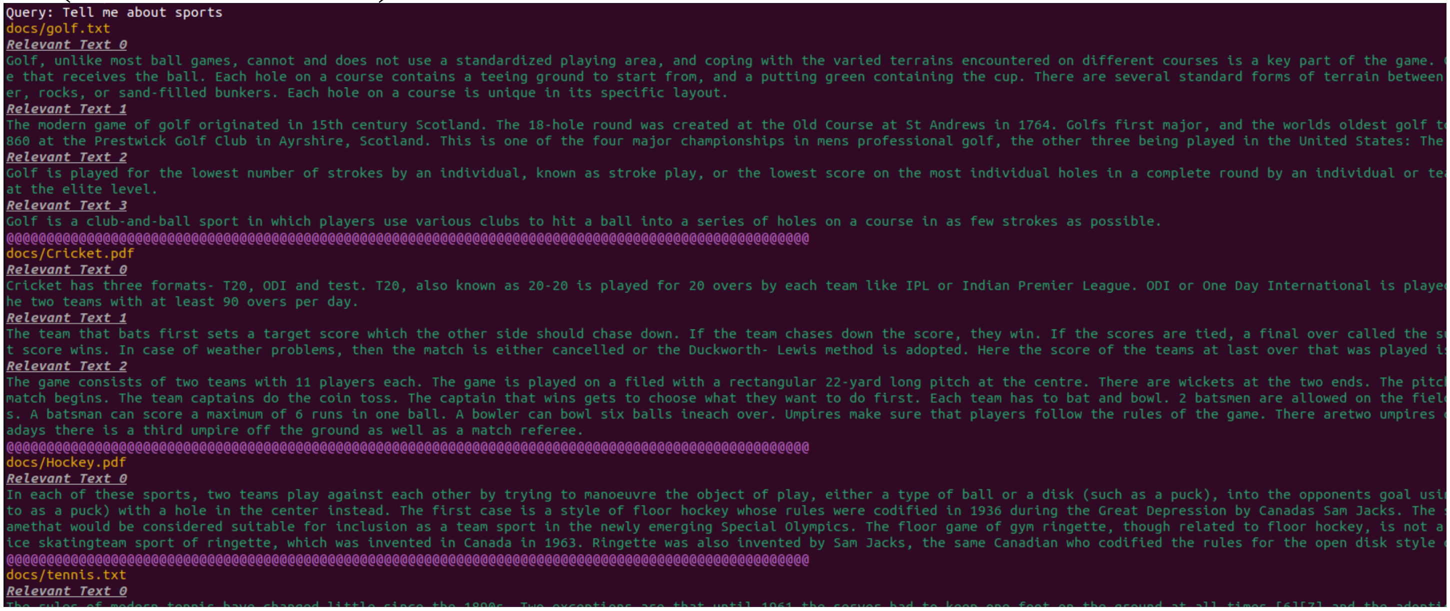
```
result_query = f"""
    SELECT scored_paragraphs.name, scored_paragraphs.paragraph, scored_paragraphs.data, scored_paragraphs.distance
    FROM (
        SELECT name, paragraph, data, Similarity(SentenceFeatureExtractor('{search_query}'), SentenceFeatureExtractor(data))
        FROM MyDocuments
    ) AS scored_paragraphs
    WHERE scored_paragraphs.distance >= {SIMILARITY_THRESHOLD}
    ORDER BY scored_paragraphs.distance DESC
    LIMIT {limit}
"""
```

The results are then ordered by similarity, and we also apply limit to the number of results. After that we output the results to the user.
6. Then, I created text summarizer using facebook/bart-large-cnn and if user needs summary, this summarizer will be called on the results returned previously.
7. Finally, the code also repeatedly checks the directory **docs** in which the documents are stored and sees if documents were added or deleted. If so, then it will call INSERT/DELETE query respectively to update MyDocuments table in which all the documents are stored.

Demo:

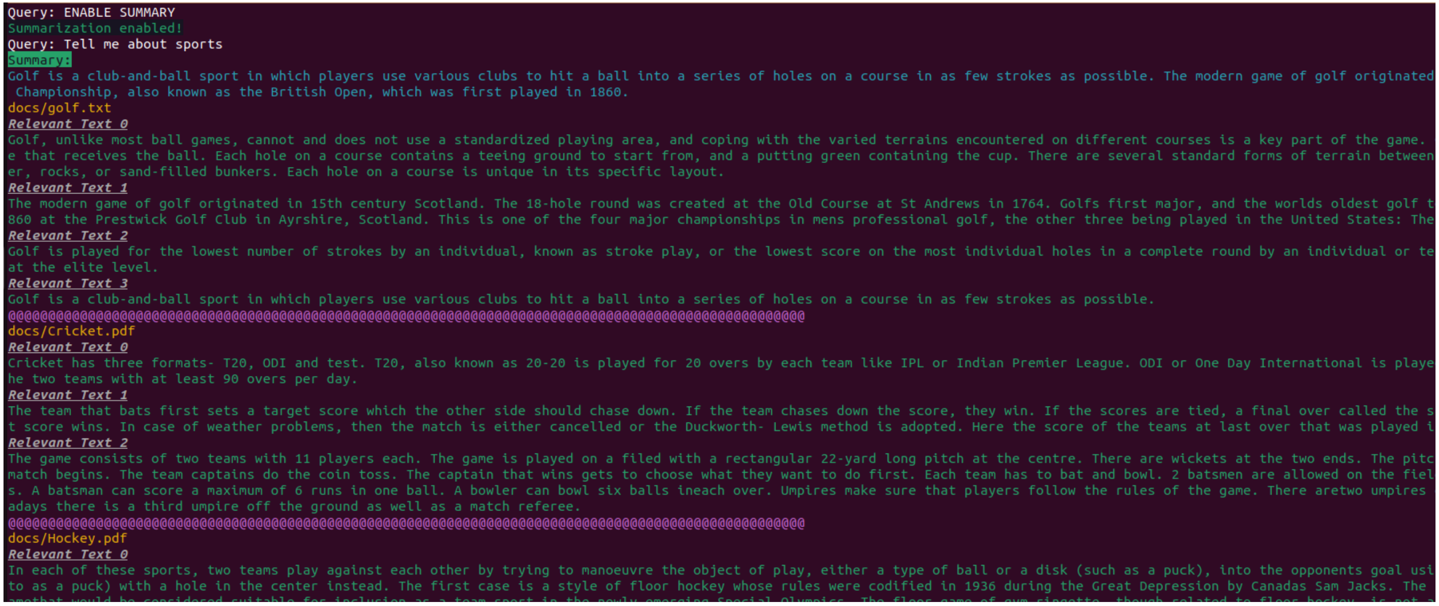
Query:

(Without Summarization)



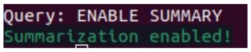
```
Query: Tell me about sports
docs/golf.txt
Relevant Text 0
Golf, unlike most ball games, cannot and does not use a standardized playing area, and coping with the varied terrains encountered on different courses is a key part of the game.
e that receives the ball. Each hole on a course contains a teeing ground to start from, and a putting green containing the cup. There are several standard forms of terrain between
er, rocks, or sand-filled bunkers. Each hole on a course is unique in its specific layout.
Relevant Text 1
The modern game of golf originated in 15th century Scotland. The 18-hole round was created at the Old Course at St Andrews in 1764. Golfs first major, and the worlds oldest golf t
860 at the Prestwick Golf Club in Ayrshire, Scotland. This is one of the four major championships in mens professional golf, the other three being played in the United States: The
Relevant Text 2
Golf is played for the lowest number of strokes by an individual, known as stroke play, or the lowest score on the most individual holes in a complete round by an individual or te
at the elite level.
Relevant Text 3
Golf is a club-and-ball sport in which players use various clubs to hit a ball into a series of holes on a course in as few strokes as possible.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
docs/Cricket.pdf
Relevant Text 0
Cricket has three formats- T20, ODI and test. T20, also known as 20-20 is played for 20 overs by each team like IPL or Indian Premier League. ODI or One Day International is playe
he two teams with at least 90 overs per day.
Relevant Text 1
The team that bats first sets a target score which the other side should chase down. If the team chases down the score, they win. If the scores are tied, a final over called the s
t score wins. In case of weather problems, then the match is either cancelled or the Duckworth- Lewis method is adopted. Here the score of the teams at last over that was played i
Relevant Text 2
The game consists of two teams with 11 players each. The game is played on a filed with a rectangular 22-yard long pitch at the centre. There are wickets at the two ends. The pitc
match begins. The team captains do the coin toss. The captain that wins gets to choose what they want to do first. Each team has to bat and bowl. 2 batsmen are allowed on the fiel
s. A batsman can score a maximum of 6 runs in one ball. A bowler can bowl six balls ineach over. Umpires make sure that players follow the rules of the game. There aretwo umpires
adays there is a third umpire off the ground as well as a match referee.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
docs/Hockey.pdf
Relevant Text 0
In each of these sports, two teams play against each other by trying to manoeuvre the object of play, either a type of ball or a disk (such as a puck), into the opponents goal usi
to as a puck) with a hole in the center instead. The first case is a style of floor hockey whose rules were codified in 1936 during the Great Depression by Canadas Sam Jacks. The
amethat would be considered suitable for inclusion as a team sport in the newly emerging Special Olympics. The floor game of gym ringette, though related to floor hockey, is not a
ice skatingteam sport of ringette, which was invented in Canada in 1963. Ringette was also invented by Sam Jacks, the same Canadian who codified the rules for the open disk style
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
docs/tennis.txt
Relevant Text 0
The rules of modern tennis have changed little since the 1890s. Two exceptions are that until 1961 the server had to keep one foot on the ground at all times [6][7] and the adopti
```

(With Summarization (Shown in Blue Text))



```
Query: ENABLE SUMMARY
Summarization enabled!
Query: Tell me about sports
Summary:
Golf is a club-and-ball sport in which players use various clubs to hit a ball into a series of holes on a course in as few strokes as possible. The modern game of golf originated
 Championship, also known as the British Open, which was first played in 1860.
docs/golf.txt
Relevant Text 0
Golf, unlike most ball games, cannot and does not use a standardized playing area, and coping with the varied terrains encountered on different courses is a key part of the game.
e that receives the ball. Each hole on a course contains a teeing ground to start from, and a putting green containing the cup. There are several standard forms of terrain between
er, rocks, or sand-filled bunkers. Each hole on a course is unique in its specific layout.
Relevant Text 1
The modern game of golf originated in 15th century Scotland. The 18-hole round was created at the Old Course at St Andrews in 1764. Golfs first major, and the worlds oldest golf t
860 at the Prestwick Golf Club in Ayrshire, Scotland. This is one of the four major championships in mens professional golf, the other three being played in the United States: The
Relevant Text 2
Golf is played for the lowest number of strokes by an individual, known as stroke play, or the lowest score on the most individual holes in a complete round by an individual or te
at the elite level.
Relevant Text 3
Golf is a club-and-ball sport in which players use various clubs to hit a ball into a series of holes on a course in as few strokes as possible.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
docs/Cricket.pdf
Relevant Text 0
Cricket has three formats- T20, ODI and test. T20, also known as 20-20 is played for 20 overs by each team like IPL or Indian Premier League. ODI or One Day International is playe
he two teams with at least 90 overs per day.
Relevant Text 1
The team that bats first sets a target score which the other side should chase down. If the team chases down the score, they win. If the scores are tied, a final over called the s
t score wins. In case of weather problems, then the match is either cancelled or the Duckworth- Lewis method is adopted. Here the score of the teams at last over that was played i
Relevant Text 2
The game consists of two teams with 11 players each. The game is played on a filed with a rectangular 22-yard long pitch at the centre. There are wickets at the two ends. The pitc
match begins. The team captains do the coin toss. The captain that wins gets to choose what they want to do first. Each team has to bat and bowl. 2 batsmen are allowed on the fiel
s. A batsman can score a maximum of 6 runs in one ball. A bowler can bowl six balls ineach over. Umpires make sure that players follow the rules of the game. There aretwo umpires
adays there is a third umpire off the ground as well as a match referee.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
docs/Hockey.pdf
Relevant Text 0
In each of these sports, two teams play against each other by trying to manoeuvre the object of play, either a type of ball or a disk (such as a puck), into the opponents goal usi
to as a puck) with a hole in the center instead. The first case is a style of floor hockey whose rules were codified in 1936 during the Great Depression by Canadas Sam Jacks. The
```

Enable Summary:

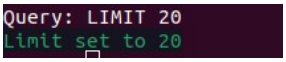


```
Query: ENABLE SUMMARY
Summarization enabled!
```

Disable Summary:

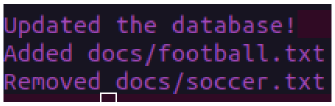
```
Query: DISABLE SUMMARY
Summarization disabled!
```

Set Limit:



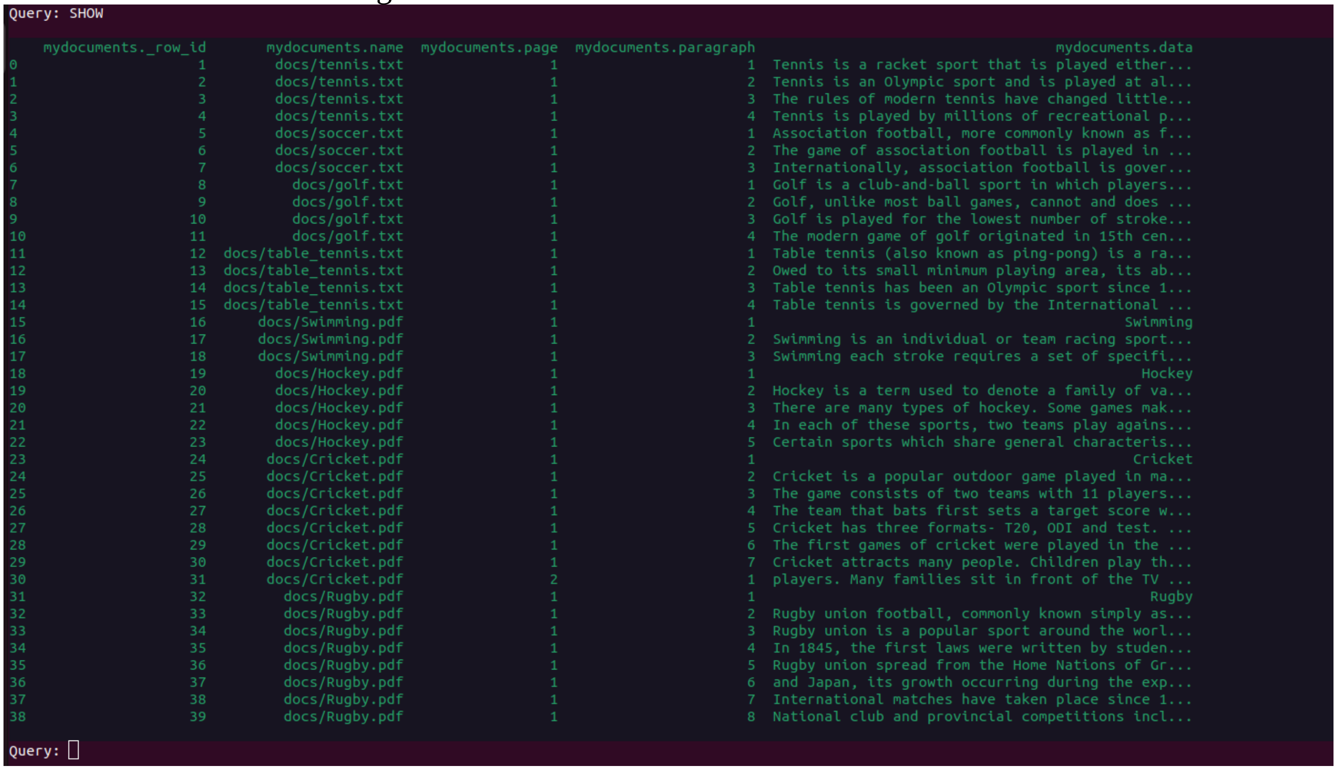
```
Query: LIMIT 20
Limit set to 20
```

Updating Database automatically:



```
Updated the database!
Added docs/football.txt
Removed docs/soccer.txt
```

SHOW Backend Table storing all documents:



```
Query: SHOW

    mydocuments._row_id       mydocuments.name  mydocuments.page  mydocuments.paragraph                                    mydocuments.data
0                     1        docs/tennis.txt                 1                      1  Tennis is a racket sport that is played either...
1                     2        docs/tennis.txt                 1                      2  Tennis is an Olympic sport and is played at al...
2                     3        docs/tennis.txt                 1                      3  The rules of modern tennis have changed little...
3                     4        docs/tennis.txt                 1                      4  Tennis is played by millions of recreational p...
4                     5        docs/soccer.txt                 1                      1  Association football, more commonly known as f...
5                     6        docs/soccer.txt                 1                      2  The game of association football is played in ...
6                     7        docs/soccer.txt                 1                      3  Internationally, association football is gover...
7                     8          docs/golf.txt                 1                      1  Golf is a club-and-ball sport in which players...
8                     9          docs/golf.txt                 1                      2  Golf, unlike most ball games, cannot and does ...
9                    10          docs/golf.txt                 1                      3  Golf is played for the lowest number of stroke...
10                   11          docs/golf.txt                 1                      4  The modern game of golf originated in 15th cen...
11                   12  docs/table_tennis.txt                 1                      1  Table tennis (also known as ping-pong) is a ra...
12                   13  docs/table_tennis.txt                 1                      2  Owed to its small minimum playing area, its ab...
13                   14  docs/table_tennis.txt                 1                      3  Table tennis has been an Olympic sport since 1...
14                   15  docs/table_tennis.txt                 1                      4  Table tennis is governed by the International ...
15                   16      docs/Swimming.pdf                 1                      1                                           Swimming
16                   17      docs/Swimming.pdf                 1                      2  Swimming is an individual or team racing sport...
17                   18      docs/Swimming.pdf                 1                      3  Swimming each stroke requires a set of specifi...
18                   19        docs/Hockey.pdf                 1                      1                                             Hockey
19                   20        docs/Hockey.pdf                 1                      2  Hockey is a term used to denote a family of va...
20                   21        docs/Hockey.pdf                 1                      3  There are many types of hockey. Some games mak...
21                   22        docs/Hockey.pdf                 1                      4  In each of these sports, two teams play agains...
22                   23        docs/Hockey.pdf                 1                      5  Certain sports which share general characteris...
23                   24       docs/Cricket.pdf                 1                      1                                            Cricket
24                   25       docs/Cricket.pdf                 1                      2  Cricket is a popular outdoor game played in ma...
25                   26       docs/Cricket.pdf                 1                      3  The game consists of two teams with 11 players...
26                   27       docs/Cricket.pdf                 1                      4  The team that bats first sets a target score w...
27                   28       docs/Cricket.pdf                 1                      5  Cricket has three formats- T20, ODI and test. ...
28                   29       docs/Cricket.pdf                 1                      6  The first games of cricket were played in the ...
29                   30       docs/Cricket.pdf                 1                      7  Cricket attracts many people. Children play th...
30                   31       docs/Cricket.pdf                 2                      1  players. Many families sit in front of the TV ...
31                   32         docs/Rugby.pdf                 1                      1                                              Rugby
32                   33         docs/Rugby.pdf                 1                      2  Rugby union football, commonly known simply as...
33                   34         docs/Rugby.pdf                 1                      3  Rugby union is a popular sport around the worl...
34                   35         docs/Rugby.pdf                 1                      4  In 1845, the first laws were written by studen...
35                   36         docs/Rugby.pdf                 1                      5  Rugby union spread from the Home Nations of Gr...
36                   37         docs/Rugby.pdf                 1                      6  and Japan, its growth occurring during the exp...
37                   38         docs/Rugby.pdf                 1                      7  International matches have taken place since 1...
38                   39         docs/Rugby.pdf                 1                      8  National club and provincial competitions incl...

Query: 
```

Exit:

```
Query: exit
Exiting...
```

Challenges:
1. Currently I have used facebook-bart-cnn summarizer and the summary results seem to omit lot of information, so I might try a different LLM for summarization or use ChatGPT.
2. Currently polling for new or deleted documents and updating database happens serially, which might block the user from entering the query if the number of documents added/deleted is quite large, hence I might try using a separate thread for this.
More discussion on ways to address the above challenges in future scope.

Future Scope:
I have thought of the following things to add to my current implementation:
1. Document clustering and more advanced queries like “Tell me about spring in context of computer science”, this will only get results from documents clustered under computer science category and give info about spring framework for Java. Or “Tell me about spring in context of nature”, this will get results from documents clustered under nature category and give info about the spring season.
2. Use ChatGPT or GPT4All to also summarize all the documents in 1 line for the user and store the summary in MyDocuments table as a new column. If user asks for summary, just out the summary of all documents. Perhaps, we can create a new abstract function for this feature.
3. I also plan to use different Sentence Feature Extractor through different models (probably from OpenAI like text-embedding-ada-002) and compare the results.
4. Currently the engine polls the directory to check for new documents sequentially every 60 seconds on the main thread. We can also do the same through a separate thread to unblock the normal flow leading to faster UI for the user.
5. Try different text summarizers and compare the results.

All of these new features can be taken up in project 2 of EvaDB.