MADRAS INSTITUTE OF TECHNOLOGY

ANNA UNIVERSITY CHENNAI

CHENNAI - 600 044.

DEPARTMENT OF INFORMATION TECHNOLOGY

IT5022 - INFORMATION RETRIEVAL

PROJECT REPORT

6/8 B. TECH INFORMATION TECHNOLOGY

E - NOTES MANAGEMENT SYSTEM

Submitted By

Prasanth S - 2022506018
Arun Kishore R - 2022506053

Introduction:

In the modern era of digital learning and knowledge sharing, efficient note organization and retrieval systems are crucial. The E-Notes Management System is a web-based platform designed to help students, educators, and researchers upload, categorize, and retrieve study notes. It enhances productivity by allowing users to store text-based content and documents and search them intelligently using Natural Language Processing (NLP) techniques.

This system is especially useful in academic institutions where large volumes of notes and documents are shared. With an intuitive interface and intelligent search capabilities, the system ensures that users can find relevant notes quickly, improving learning outcomes and reducing time spent on manual searches.

Objectives:

- To create a centralized platform for uploading, managing, and accessing academic notes.
- To provide a simple and secure login system for students and faculty.
- To enable easy search and retrieval of notes using keywords or categories.
- To ensure quick access to notes from any device, anytime.
- To reduce the use of paper by digitizing academic materials.

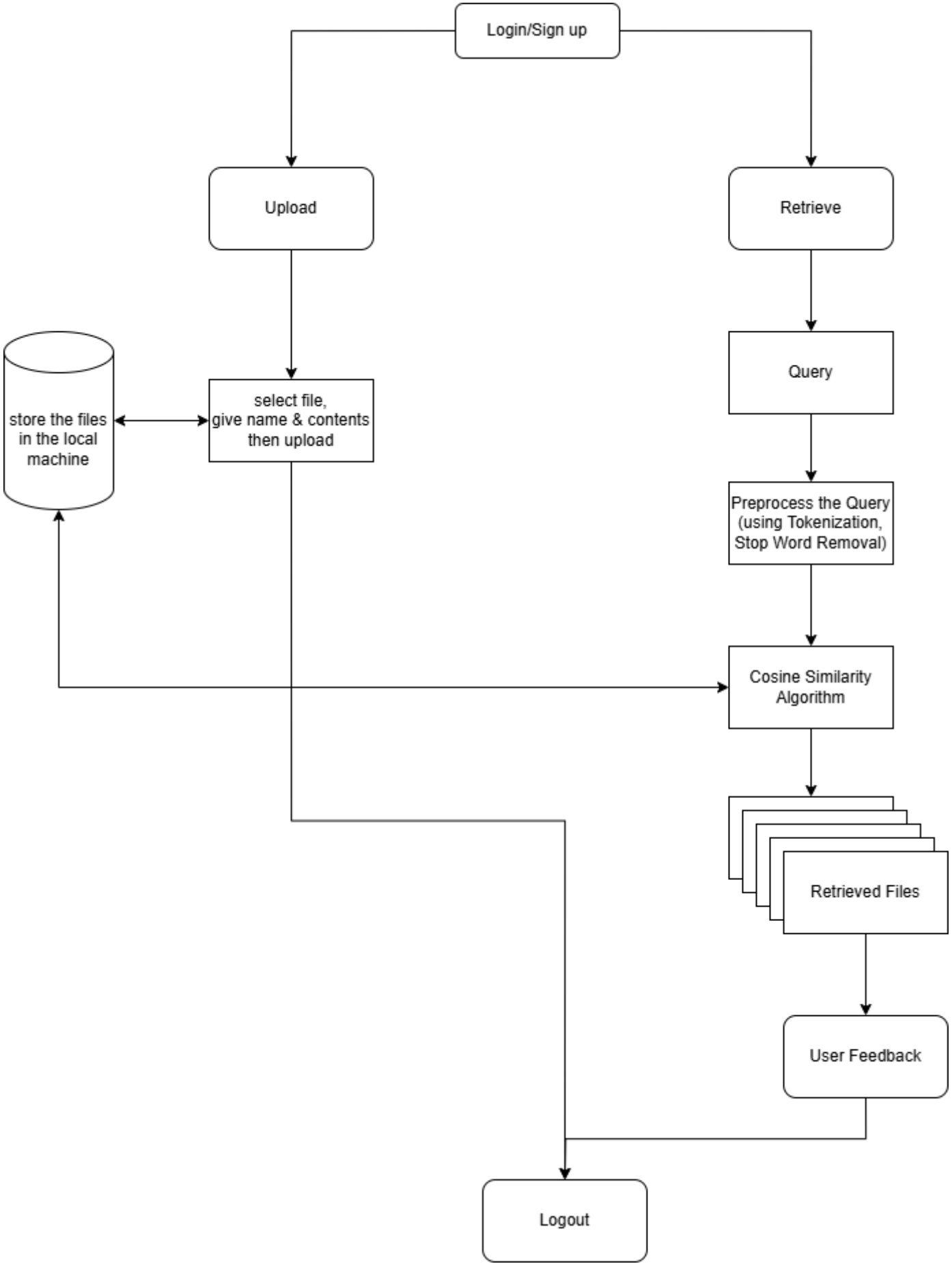
Explanation:

The E-Notes Management System allows authenticated users to:

- Register and log in securely
- Upload notes with titles, categories, and optional file attachments (PDFs, images, documents)
- Browse all notes in the repository
- Search notes based on queries using advanced Information Retrieval techniques
- View detailed note content and provide feedback on search effectiveness

The core strength of this system lies in its intelligent search feature. Unlike basic keyword matching, it employs TF-IDF vectorization and Cosine Similarity to return results that are semantically relevant, even if the exact words don't match.

Flow Diagram:



Technologies Used:

Frontend:

- HTML/CSS for basic UI layout
- Jinja2 for dynamic content rendering
- Bootstrap (optional) for responsive design

Backend:

- Python with Flask as the web framework
- Flask-Login for user session management
- MySQL for lightweight data storage

Machine Learning/NLP:

- NLTK for tokenization, lemmatization, and text preprocessing
- Scikit-learn for TF-IDF Vectorization and Cosine Similarity calculations

Advantages:

- Saves developer time in writing comments manually.
- Enhances understanding of complex codebases.
- Useful in educational tools and documentation generators.

How Information Retrieval Works in this E Notes Management System:

The intelligent search functionality is powered by a basic Information Retrieval pipeline, which includes:

1. Text Preprocessing (Normalization)
Purpose: To clean and prepare both the user's search query and the note contents for better comparison.
Techniques Used:

- Converting text to lowercase.
- Removing punctuation using str.translate.
- Tokenizing the text with nltk.word_tokenize.
- Removing English stopwords using NLTK's stopwords.words.
- Applying part-of-speech tagging using nltk.pos_tag.
- Lemmatizing tokens based on POS tags using WordNetLemmatizer.

This normalization ensures that all texts are reduced to their meaningful core forms, improving the accuracy of later vectorization and matching.

2. Feature Extraction using TF-IDF

Purpose: To convert textual data into a numerical representation for similarity calculation.

Implementation:

- The system uses TfidfVectorizer from Scikit-learn to compute TF-IDF vectors for each note and the user's query.
- Both the note contents and the search query go through the same preprocessing pipeline before vectorization.

This step helps quantify the importance of terms in documents relative to the entire corpus.

3. Similarity Calculation

Purpose: To identify and rank notes that are most relevant to the user's query.

Implementation:

- Cosine similarity is computed between the TF-IDF vector of the search query and each note.
- Each note is assigned a similarity score.
- Additional relevance boosting is done manually by checking if any note words start with query terms or if the title starts with a query term.

This similarity score determines how closely a note matches the query and is used to rank the results.

4. Query Expansion (Synonym Enrichment)

Purpose: To improve recall by finding more relevant documents even if the exact terms do not match.

Implementation:

- WordNet is used to find synonyms for each query term using wordnet.synsets.
- Lemmas from the synsets are extracted and added to the list of query terms.
- These expanded terms are then used for matching and highlighting.

This allows the system to recognize conceptually similar content beyond exact keyword matches.

5. Highlighting Relevant Content

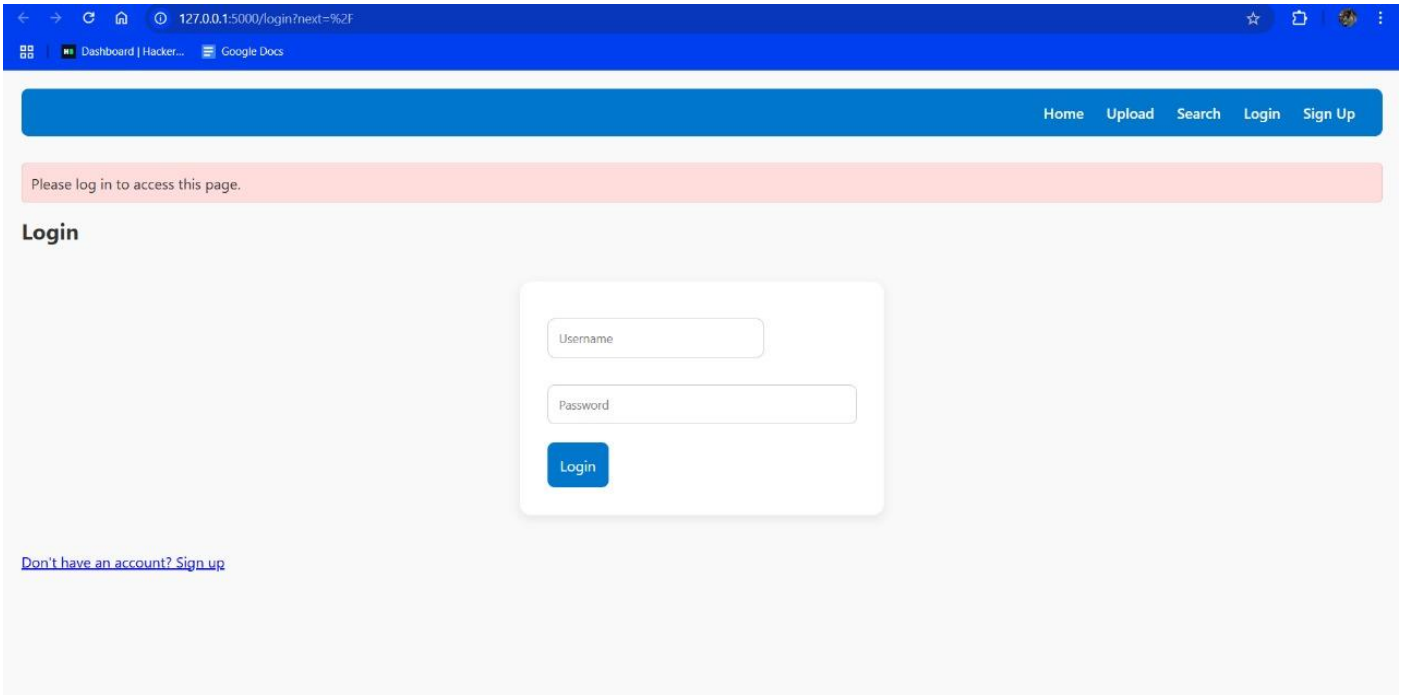
Purpose: To visually emphasize matched keywords in search results.

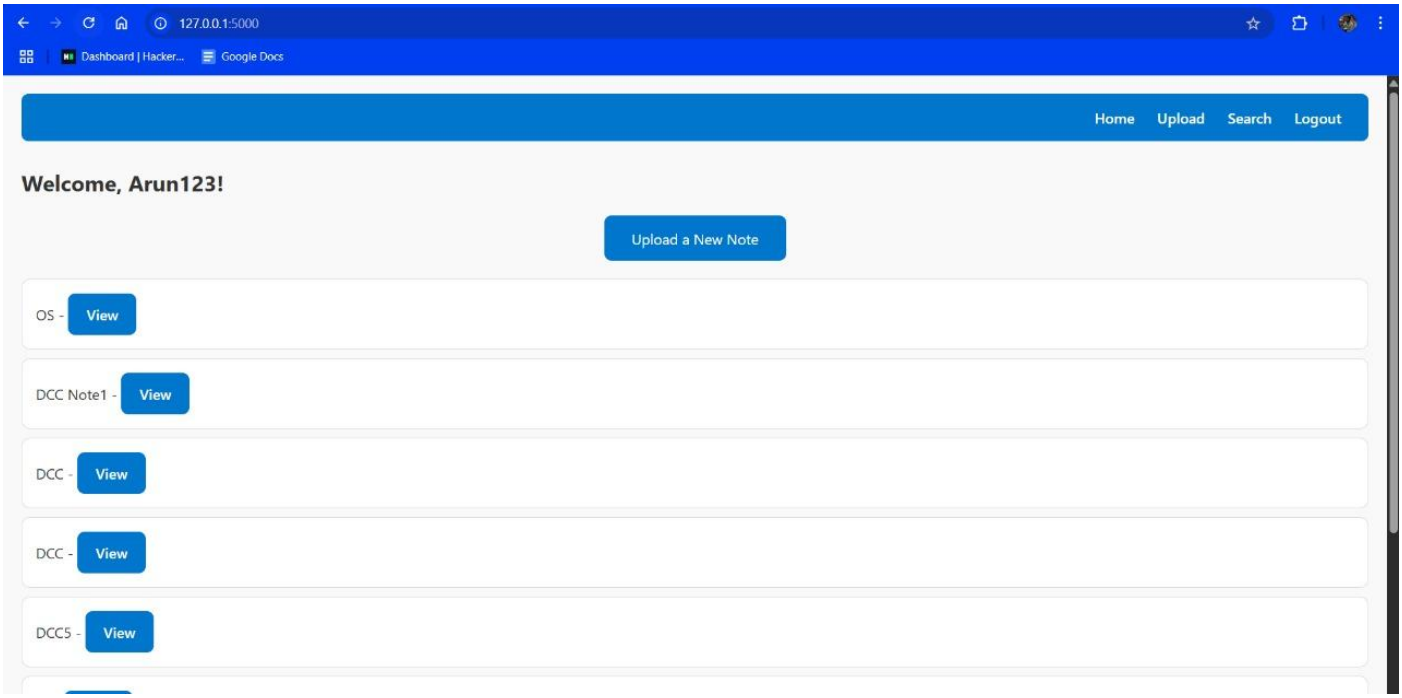
Implementation:

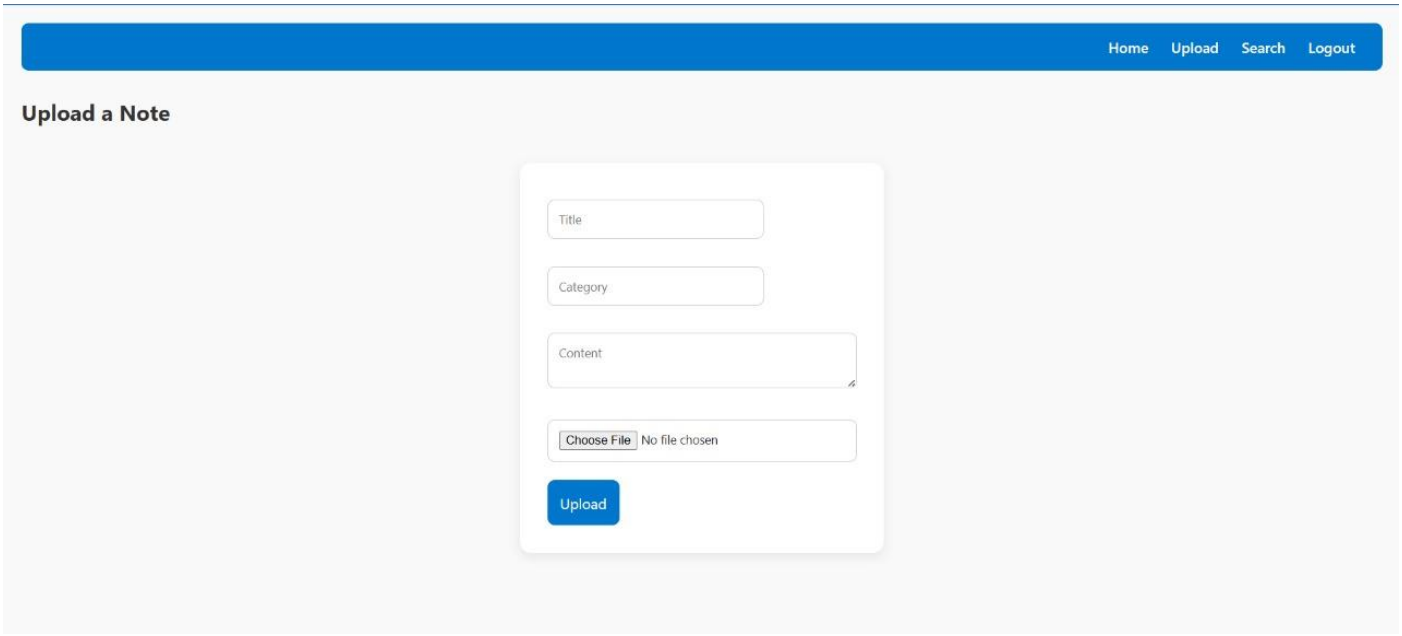
- Uses regular expressions to highlight matching terms from the expanded query in the note title and content.
- Highlighted terms are wrapped in <mark> tags for rendering in the template.

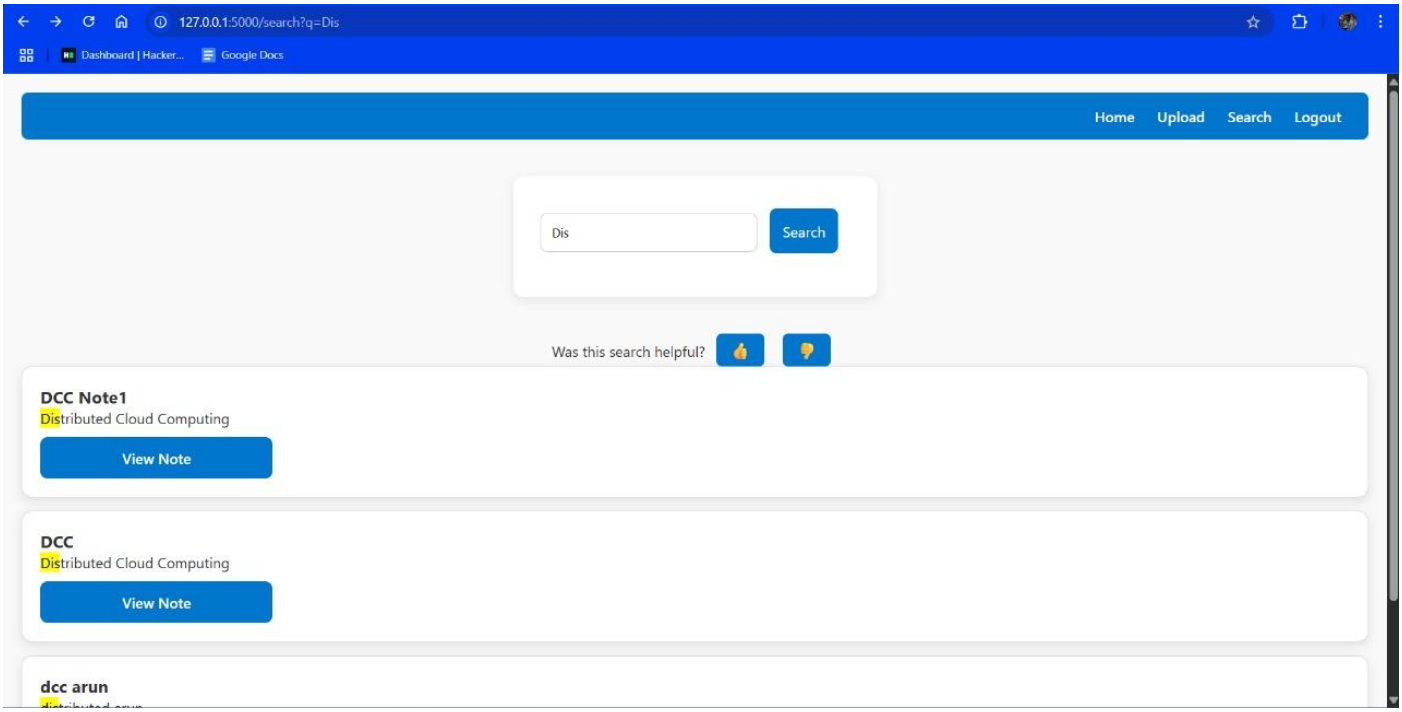
This enhances user experience by making the relevance of results clearer.

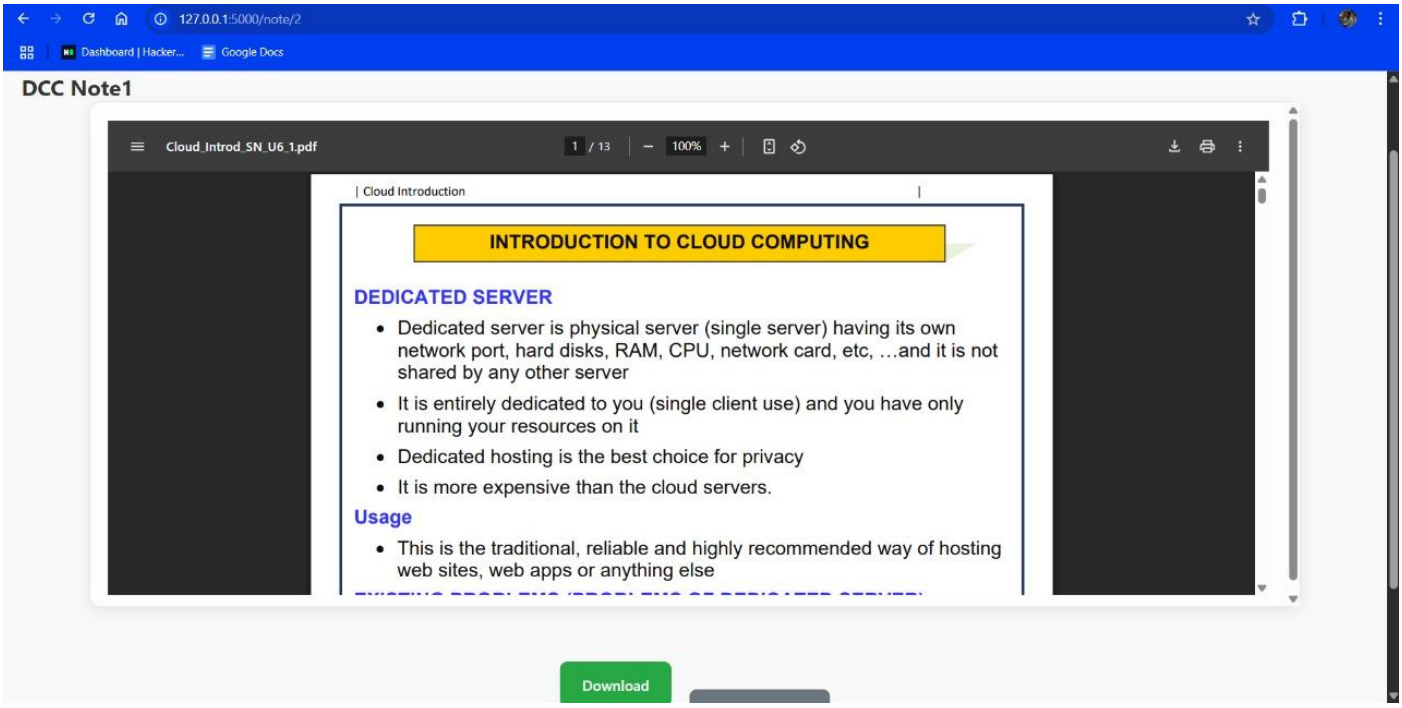
OUTPUT:


127.0.0.1:5000/login?next=%2F
Dashboard | Hacker...
Google Docs
Home
Upload
Search
Login
Sign Up
Please log in to access this page.
Login
Username
Password
Login
Don't have an account? Sign up


127.0.0.1:5000
Dashboard | Hacker...
Google Docs
Home
Upload
Search
Logout
Welcome, Arun123!
Upload a New Note
OS - View
DCC Note1 - View
DCC - View
DCC - View
DCC5 - View


Home
Upload
Search
Logout
Upload a Note
Title
Category
Content
Choose File
No file chosen
Upload


127.0.0.1:5000/search?q=Dis
Dashboard | Hacker...
Google Docs
Home
Upload
Search
Logout
Dis
Search
Was this search helpful?
DCC Note1
Distributed Cloud Computing
View Note
DCC
Distributed Cloud Computing
View Note
dcc arun



Conclusion:

The E-Notes Management System provides an effective solution for managing and retrieving academic content in a centralized, intelligent manner. By leveraging natural language processing techniques like TF-IDF and Cosine Similarity, the system goes beyond simple keyword matching to provide meaningful search results. With a robust backend, an intuitive interface, and modular architecture, this system has strong potential for further development and real-world academic deployment.