SMART INDIA HACKATHON 2022

Basic Details of the Team and Problem Statement

Ministry/Organization Name/Student Innovation: Ministry of Ayurveda, Yoga, Naturopathy, Unani, Siddha, Sowa-Rigpa and Homoeopathy (AYUSH)

PS Code: DK734

Problem Statement Title: Centralized thesis repository (MD/PHD) Repository

Team Name: Samadrishya

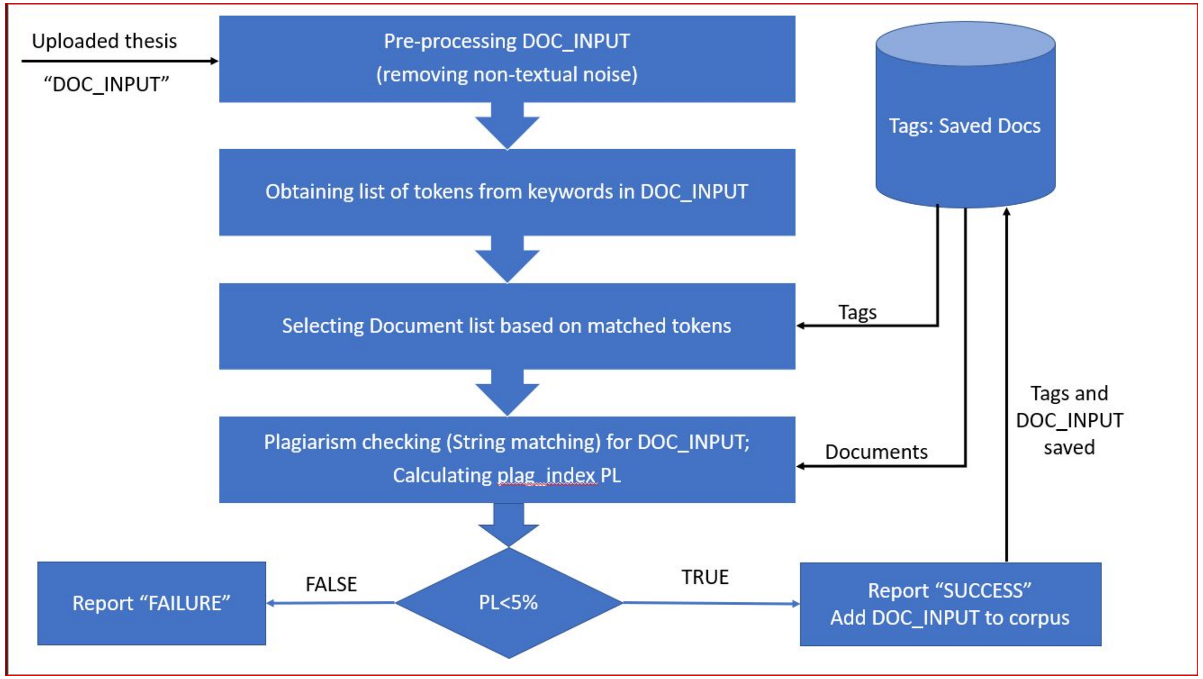
Team Leader Name: Mupudi Monica

Institute Code (AISHE): 424

Institute Name: IIIT Bhubaneswar

Theme Name: Smart Education (Software)

Idea/Approach Details


Uploaded thesis
"DOC_INPUT"
Pre-processing DOC_INPUT (removing non-textual noise)
Obtaining list of tokens from keywords in DOC_INPUT
Selecting Document list based on matched tokens
Plagiarism checking (String matching) for DOC_INPUT; Calculating plag_index PL
Tags: Saved Docs
Tags
Documents
Tags and DOC_INPUT saved
Report "FAILURE"
FALSE
PL<5%
TRUE
Report "SUCCESS" Add DOC_INPUT to corpus

Idea/Solution/Prototype :

1. Uploading the thesis
2. Image Pre-processing on the thesis
3. Obtaining list of tokens from the keywords in the thesis by word-processing
4. Matching the tokens from tags in the database using Rabin–Karp algorithm with Hash
5. Calculation of Plagiarism_index PL
6. Acceptance/Rejection based on PL

Admins can:

1. View all the accepted Documents
2. Find the plagiarised phrases present in the accepted Document across different files.

Technology stack :

Frontend: HTML5, CSS3, ReactJS

Backend: Python, including modules:

- PyPDF (for pdf handling)
- OpenCV2 (image preprocessing)
- PyTesseract (Text recognition)
- Mongoose (for MongoDB)

Database: MongoDB

Idea/Approach Details

Use Cases:

- **ADMIN:** Login to
  - View / Download documents in corpus.
  - Obtain the list of plagiarised phrases between two documents.
- **FREE USERS** (no registration required)**:**
  - Upload thesis and obtain PL index.
  - Compare thesis against tagged documents in the corpus and find plagiarized phrases.
  - Receive confirmation/rejection of upload.
  - View/download documents in corpus.

Dependencies / Limitations:

- **Time Complexity:** String comparing algorithms are computationally extensive. However, using tags based on theses keywords reduces the document domain to compare against, improving performance times.
- **Including online thesis repositories:** Currently, incoming thesis are compared to the documents present in the corpus. Including other online repositories will enhance unique presentations.
- **Expansion to other file formats:** With minor changes, the project can be expanded for epub and docx formats.

Team Member Details

Team Leader Name: Mupudi Monica (B119034)

Branch : B. Tech   Stream : CSE   Year : III Year

Team Member 1 Name: Atrik Ray (B219015)

Branch : B. Tech   Stream : CSE   Year : III Year

Team Member 2 Name: Manas Sahu (B219032)

Branch : B. Tech   Stream : CSE   Year : III Year

Team Member 3 Name: Abhishek Jaiswal (B419007)

Branch : B. Tech   Stream : CSE   Year : III Year

Team Member 4 Name: Gayathri M S (B219024)

Branch : B. Tech   Stream : CSE   Year : III Year

Team Member 5 Name: Nikhil Kumar Patra (B519030)

Branch : B. Tech   Stream : CE   Year : III Year