A MACHINE LEARNING–BASED SEMANTIC TEXT PLAGIARISM DETECTION SYSTEM

University of Information Technology, HCMC, Viet Nam
Vietnam National University, HCMC, Viet Nam

What ?

A machine learning–based system for detecting plagiarism in textual documents.

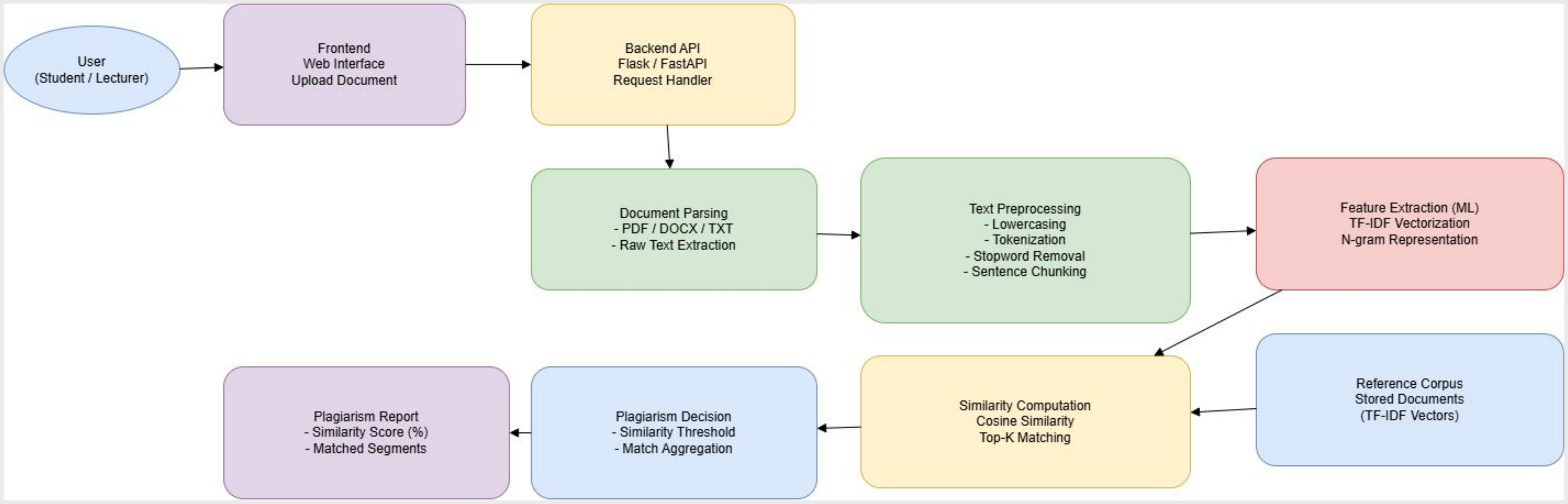
The system analyzes semantic similarity between documents using text representations and similarity measures to identify copied or rephrased content.

Why ?

Traditional plagiarism detection methods mainly rely on keyword or exact string matching, which fail to detect paraphrase plagiarism.

Machine learning–based semantic representations enable more effective and meaningful plagiarism detection.

Overview



User
(Student / Lecturer)

Frontend
Web Interface
Upload Document

Backend API
Flask / FastAPI
Request Handler

Document Parsing
- PDF / DOCX / TXT
- Raw Text Extraction

Text Preprocessing
- Lowercasing
- Tokenization
- Stopword Removal
- Sentence Chunking

Feature Extraction (ML)
TF-IDF Vectorization
N-gram Representation

Reference Corpus
Stored Documents
(TF-IDF Vectors)

Similarity Computation
Cosine Similarity
Top-K Matching

Plagiarism Decision
- Similarity Threshold
- Match Aggregation

Plagiarism Report
- Similarity Score (%)
- Matched Segments

Description

1. Problem Formulation

- Plagiarism detection is formulated as a semantic similarity problem between an input document and a reference corpus.
- Rather than relying on exact word overlap, the task focuses on identifying content reuse at the meaning level, including paraphrasing and structural rewriting.
- This formulation enables the system to detect plagiarism beyond surface lexical similarity.

2. System Architecture

- The system follows a multi-stage pipeline consisting of preprocessing, semantic representation, and similarity analysis.
- Input documents are first cleaned, normalized, and segmented into smaller textual units.
- Each segment is transformed into a semantic vector using machine learning–based text representations such as TF-IDF and n-gram features.
- Similarity computation is then performed in the vector space to identify highly similar text segments between documents.

3. Output and Evaluation

- The system produces a plagiarism report containing similarity scores and detected suspicious segments.
- These results provide transparent and interpretable evidence to support plagiarism assessment.
- The system is designed to be practical, scalable, and suitable for real-world academic use.

Nguyen Trong Tuan Anh – University of Information Technology, HCMC, Viet Nam
TEL : 0975824272 Email : anhngtt.20@grad.uit.edu.vn