

HỆ THỐNG PHÁT HIỆN ĐẠO VĂN VĂN BẢN DỰA TRÊN BIỂU DIỄN NGỮ NGHĨA VÀ MACHINE LEARNING

Nguyễn Trọng Tuấn Anh - 250101004

Tóm tắt

- Lớp: CS2205.CH201
- Link Github: [tuananhdev1102/CS2205.CH201](https://github.com/tuananhdev1102/CS2205.CH201)
- Link YouTube: [CS2205 CH201](https://www.youtube.com/watch?v=CS2205CH201)
- Nguyễn Trọng Tuấn Anh - 250101004



Giới thiệu

Thực trạng:

- Gia tăng bài nộp điện tử -> Gia tăng nguy cơ gian lận học thuật.
- Đạo văn tinh vi (diễn đạt lại, thay từ đồng nghĩa) ngày càng phổ biến.

Hạn chế của phương pháp truyền thống:

- Dựa trên so khớp từ khóa/chuỗi ký tự.
- Chỉ phát hiện được sao chép nguyên văn, bỏ qua các trường hợp sao chép ý tưởng.

Nhu cầu: Cần một hệ thống thông minh hơn, phát hiện đạo văn ở cấp độ ý nghĩa (ngữ nghĩa).

Mục tiêu

Xây dựng Pipeline Xử Lý:

- Thiết lập quy trình tiền xử lý, phân đoạn văn bản.
- Ứng dụng mô hình ML có sẵn để chuyển văn bản thành vector ngữ nghĩa.

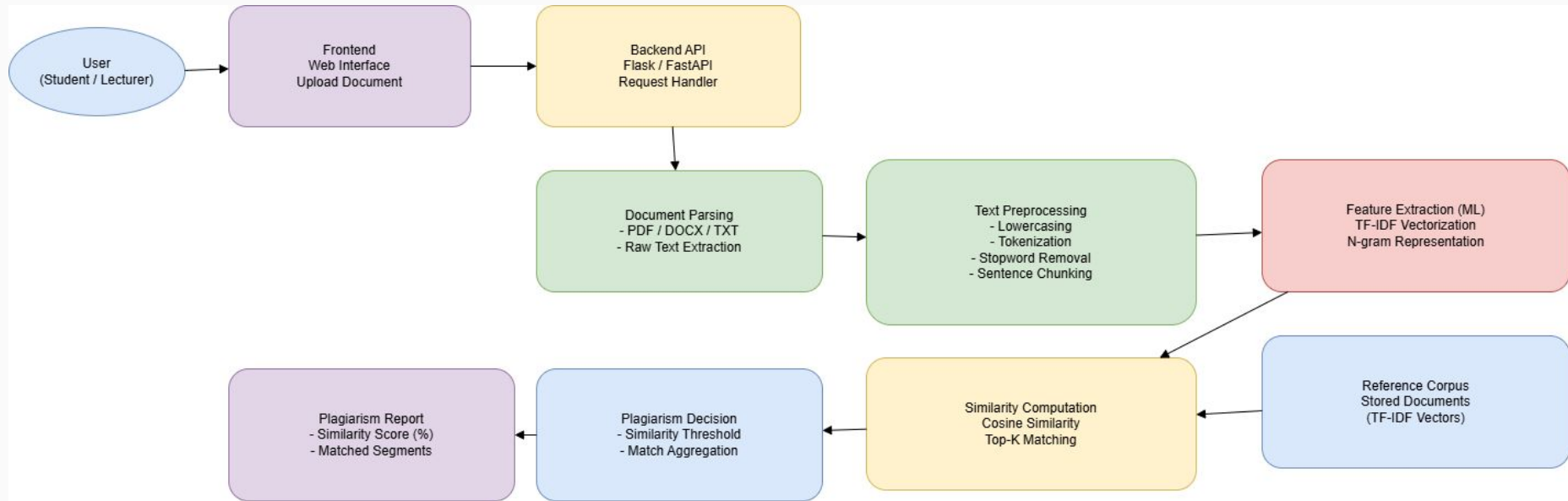
Phát triển Hệ Thống So Sánh Ngữ Nghĩa:

- Thiết kế hệ thống tính toán độ tương đồng giữa văn bản kiểm tra và tài liệu tham khảo.
- Nhận diện được cả sao chép nguyên văn và diễn đạt lại.

Đánh giá & Phân Tích Tính Khả Thi:

- Kiểm tra hiệu suất bằng các số liệu định lượng.
- Đánh giá tiềm năng ứng dụng thực tế trong môi trường giáo dục.

Nội dung và Phương pháp



Sơ đồ pipeline tổng quát của hệ thống phát hiện đạo văn văn bản dựa trên Machine Learning

Nội dung và Phương pháp

- Tiếp nhận tài liệu từ người dùng
- Trích xuất văn bản từ PDF / DOCX
- Tiền xử lý
- Trích xuất đặc trưng: TF-IDF, N-gram
- Tính độ tương đồng: Cosine Similarity
- Phân tích & tổng hợp kết quả

Kết quả dự kiến

Một hệ thống hoạt động được:

- Giao diện đơn giản để nhập văn bản và nhận kết quả.
- Khả năng xử lý ổn định, rõ ràng.

Báo cáo kết quả chi tiết và trực quan:

- Điểm số tương đồng tổng thể và từng phần.
- Highlight các đoạn văn khả nghi, chỉ ra nguồn tham khảo có khả năng trùng lặp.
- Biểu đồ, hình ảnh hóa mối quan hệ tương đồng.

Đánh giá hiệu quả: Bảng so sánh khả năng phát hiện giữa sao chép trực tiếp và diễn đạt lại.

Tài liệu tham khảo

- [1]. Christopher D. Manning, Hinrich Schütze:
Foundations of Statistical Natural Language Processing. MIT Press, 1999.
- [2]. Dan Jurafsky, James H. Martin:
Speech and Language Processing. Pearson, 3rd Edition, 2025.
- [3]. Sebastian Raschka, Vahid Mirjalili:
Python Machine Learning. Packt Publishing, 3rd Edition, 2019.
- [4]. Wikipedia:
Plagiarism Detection. URL: https://en.wikipedia.org/wiki/Plagiarism_detection
(truy cập 30-12-2025).