PBL 6: **Hệ thống học từ vựng tiếng Anh thông minh tích hợp AI tạo Flashcard từ PDF**

1. **Tên đề tài**

Hệ thống học từ vựng tiếng Anh thông minh tích hợp AI tạo Flashcard từ PDF.

1. **Mục tiêu**

* Xây dựng một ứng dụng web hỗ trợ người học tiếng Anh tự động tạo flashcard từ file PDF (giáo trình, bài báo, tài liệu học tập).
* Ứng dụng giúp người học tiết kiệm thời gian trong việc ghi chú từ vựng, tăng cường khả năng ghi nhớ bằng phương pháp flashcard.

Mục tiêu cần đạt được:

* Xử lý tài liệu PDF: cho phép người dùng upload PDF và tự động trích xuất văn bản từ file PDF.
* Xử lý ngôn ngữ tự nhiên (NLP): Lọc danh sách từ vựng xuất hiện trong tài liệu rồi phân loại từ phổ biến/dễ với ít gặp (dựa trên dataset). Tạo danh sách “từ mới cần học”.
* Tạo flashcard tự động: flashcard bao gồm từ vựng, loại từ, nghĩa tiếng Việt, phiên âm, ví dụ minh họa, phát âm (audio). Cho phép người học lật flashcard để ghi nhớ.
* Hệ thống web: Giao diện đơn giản, thân thiện với người dùng.
* Chức năng cơ bản: hoàn thành được các chức năng cơ bản của web.
* Mô hình tạo trắc nghiệm

1. **Các chức năng chính**

* Đăng nhập, đăng kí.
* Tạo flashcard thủ công.
* Upload PDF lọc từ mới, phân loại mức độ từ rồi tạo flashcard.
* Học và ôn tập flashcard (làm bài test).
* Chia sẻ flashcard.
* Thống kê quá trình học.
* Tạo Group để giao lưu, học hỏi.

1. **Học máy áp dụng**

* CNN (Convolutional Neural Network – Character-level): dự đoán độ khó của từ vựng dựa trên đặc trưng ký tự (số ký tự, cấu trúc, mẫu chữ cái).
* Logistic Regression: dự đoán xác suất quên từ trong hệ thống Spaced Repetition, dựa trên:
  + số lần ôn đã học
  + khoảng thời gian từ lần học gần nhất
  + độ khó của từ

1. **Quy trình học máy**
   1. Tiền xử lý dữ liệu

* Lấy dữ liệu văn bản từ PDF
* Tách từ
* Xác định từ mới (so sánh với từ đã biết)
  1. Gán nhãn dữ liệu
* Độ khó từ vựng: dựa theo tần suất xuất hiện, độ dài từ, mức độ hiếm.
* Lịch sử ôn tập: số lần ôn, thời gian cách lần trước, trạng thái nhớ/quên.
  1. Huấn luyện mô hình
* CNN: đầu vào là từ dạng ký tự, đầu ra là mức độ khó.
* Logistic Regression: đầu vào là (số lần ôn, khoảng cách thời gian, độ khó), đầu ra là xác suất quên.
  1. Dự đoán & Ứng dụng
* Tự động gợi ý flashcard với từ mới + độ khó.
* Dự đoán từ nào sắp quên để ưu tiên ôn lại.
* Sinh lịch ôn tập cá nhân hóa dựa trên dự đoán trên để thông báo.

1. **Quản trị mạng và hệ thống**
   1. Kiến trúc hệ thống

Hệ thống được thiết kế theo mô hình client-server:

* Client Apps (Web): Giao diện người dùng, nơi học viên học từ mới và ôn tập.
* Load Balancer: Phân phối đều request từ client đến các server, đảm bảo hệ thống ổn định khi nhiều người truy cập.
* API Server (Spring Boot + Tomcat): Xử lý logic ứng dụng, kết nối giữa client và các dịch vụ phía sau.
* ML Model Service: Triển khai mô hình học máy (lọc từ mới, dự đoán quên từ, gợi ý lịch ôn tập).
* Database

Luồng hoạt động:

Client → Load Balancer → API Server → ML Model / Database → Trả kết quả về Client.

* 1. Cấu hình server
* Hệ điều hành: Ubuntu
* Web Server: Nginx (reverse proxy, phục vụ file tĩnh) + Spring Boot (Tomcat nhúng)
* Cơ sở dữ liệu: MySQL (lưu trữ dữ liệu người dùng) + Redis (bộ nhớ đệm tăng tốc)
* ML Serving: TensorFlow Serving hoặc FastAPI để triển khai mô hình học máy

1. **Kiểm thử phần mềm**

Để đảm bảo chất lượng hệ thống, nhóm sử dụng các công cụ sau:

* Kiểm thử API
  + **Postman**: Gửi request đến API, kiểm tra tính đúng đắn và hiệu năng của các endpoint.
  + **JUnit + Spring Boot Test**: Viết test case tự động cho các service trong backend.
* Kiểm thử hiệu năng & tải
  + **Apache JMeter**: Mô phỏng nhiều người dùng truy cập đồng thời, đo độ trễ, throughput, khả năng chịu tải của hệ thống.
* Kiểm thử giao diện
  + **Selenium**: Tự động hóa kiểm thử giao diện web, kiểm tra tính đúng đắn khi người dùng thao tác.

1. **Công nghệ sử dụng**
   1. **Frontend**: ReactJS – xây dựng giao diện người dùng.
   2. **Backend**: Java Spring Boot – xử lý logic nghiệp vụ, xây dựng RESTful API.
   3. **Database**: MySQL – lưu trữ dữ liệu.
   4. **Công cụ hỗ trợ**: VSCode, Git/GitHub – viết code, quản lý và lưu trữ mã nguồn.
2. **Kết quả kì vọng**

* Web cho phép người dùng upload PDF → tạo flashcard tự động.
* Flashcard có đầy đủ từ, nghĩa, phát âm, ví dụ.
* Hệ thống gợi ý lịch ôn tập cá nhân hóa.
* Thống kê quá trình học.
* Web cho phép người dùng tạo Group học chung