1. Seminar Title

Doc Parsing: So sánh các Công cụ Truyền thống (pdfplumber) và Hiện đại (Dolphin)

2. Introduction / Background

Trong phát triển ứng dụng web hiện đại, nhu cầu xử lý và trích xuất thông tin từ các tài liệu phi cấu trúc như PDF, hình ảnh, hay file DOCX ngày càng tăng. Các tài liệu này thường được thiết kế cho con người đọc (chứ không thuận tiện cho máy tính đọc), khiến việc trích xuất dữ liệu tự động trở nên phức tạp. Việc phân tích tài liệu một cách chính xác là bước đầu tiên và quan trọng trong nhiều quy trình nghiệp vụ, từ nhập hóa đơn, xử lý CV, đến việc xây dựng cơ sở tri thức cho các hệ thống AI (như RAG). Seminar này sẽ khám phá các thách thức và so sánh những công cụ giúp giải quyết vấn đề này.

3. Objectives

- Phân tích các thách thức chính khi phân tích tài liệu (ví dụ: bố cục phức tạp, văn bản trong hình ảnh, bảng biểu bị xáo trộn).
- So sánh hai phương pháp tiếp cận chính: phương pháp truyền thống dựa trên tọa độ (ví dụ: pdfplumber) và phương pháp hiện đại dựa trên thị giác máy tính và AI (ví dụ: Dolphin).
- Trình diễn (demo) cơ bản cách sử dụng các công cụ này để trích xuất văn bản và bảng biểu từ các loại PDF khác nhau.

4. Scope and Relevance

- **Phạm vi (Scope):** Hội thảo sẽ tập trung vào việc trích xuất văn bản và bảng biểu từ các tệp PDF—bao gồm cả PDF gốc (digitally native) và PDF được quét (scanned). Chúng ta sẽ so sánh trực tiếp pdfplumber với các khả năng của mô hình AI như Dolphin và Unstructured.io. Hội thảo sẽ không đi sâu vào việc huấn luyện các mô hình AI, mà tập trung vào việc ứng dụng các công cụ có sẵn.
- **Mức độ liên quan (Relevance):** Đối với các nhà phát triển web, việc nắm vững các công cụ này cho phép họ xây dựng các tính năng có giá trị cao, ví dụ: tự động hóa việc nhập dữ liệu từ hóa đơn, chuẩn hóa CV ứng viên, hoặc trích xuất nội dung từ tài liệu học thuật cho các ứng dụng EdTech. Đối với doanh nghiệp, tự động hóa quy trình này giúp tiết kiệm đáng kể thời gian, chi phí và giảm thiểu lỗi sai so với việc nhập liệu thủ công.

5. Expected Outcomes

Sau khi tham dự, khán giả sẽ: