ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA CÔNG NGHỆ PHẦN MỀM

BÁO CÁO

CHUẨN BỊ DỮ LIỆU CHO HỆ THỐNG UIT AI ASSISTANT

NHÓM THỰC HIỆN:
Quách Gia Kiệt - 23520819
Nguyễn Tuấn Kiệt - 23520815

GV HƯỚNG DẪN:
Th.S Nguyễn Công Hoan

TP. HỒ CHÍ MINH, 2025

Mục lục

Chương 1 Giới thiệu

Hệ thống UIT AI Assistant sử dụng kỹ thuật RAG (Retrieval-Augmented Generation) để trả lời câu hỏi của sinh viên về quy định đào tạo và chương trình học. Dữ liệu được thu thập từ các nguồn chính thức của trường, bao gồm:

- Website DAA.UIT: Chương trình đào tạo các khóa
- Website chính thức UIT: Quyết định, thông báo, quy chế

Quy mô dữ liệu:

Danh mục	**Số lượng**	**Dung lượng**	**Định dạng**
Quy định (regulation)	27 file	∼30 MB	PDF
Chương trình đào tạo (curriculum)	100+ file	∼69,000 dòng	Markdown

Bảng 1.1: Thống kê dữ liệu thô

Chương 2 Cấu trúc dữ liệu

2.1 Dữ liệu quy định (regulation)

Các file PDF quy định có cấu trúc văn bản hành chính chuẩn:

- **Header:** Logo, tên trường, số quyết định, ngày ban hành
- **Body:** Phân cấp theo CHƯƠNG > Điều > Khoản > Mục (a, b, c)
- **Bảng biểu:** Danh sách môn học, điểm số, học phí
- **Footer:** Chữ ký, phụ lục, số trang

Ví dụ: Quyết định 790 - Quy chế đào tạo (28/9/2022) có 27 trang, 5 chương, 40+ điều về tổ chức đào tạo, đăng ký học, thi cử, xét tốt nghiệp.

2.2 Dữ liệu chương trình đào tạo (curriculum)

Các file Markdown có cấu trúc:

- **Phần 1:** Mục tiêu đào tạo, vị trí việc làm, quan điểm xây dựng chương trình
- **Phần 2:** Chuẩn đầu ra (LO1-LO10)
- **Phần 3:** Khối kiến thức
  - Đại cương: 45 TC (lý luận chính trị, toán-tin, ngoại ngữ)
  - Chuyên nghiệp: 68 TC (cơ sở nhóm ngành, cơ sở ngành, chuyên ngành)
  - Tốt nghiệp: 12 TC (đồ án, khóa luận)
- **Phần 4:** Danh sách môn học chi tiết (Mã môn, Tên môn, TC, LT, TH)
- **Phần 5:** Điều kiện tốt nghiệp

Ví dụ: Chương trình Cử nhân CNTT khóa 2020 có $\geq$125 TC, chia thành 4 hướng chuyên ngành (phân tích dữ liệu, quản lý doanh nghiệp, web, ứng dụng CNTT).

Chương 3 Đặc trưng dữ liệu

3.1 Đặc điểm chung

- **Ngôn ngữ:** Tiếng Việt, có thuật ngữ tiếng Anh
- **Văn phong:** Hành chính (quy định), học thuật (chương trình)
- **Cấu trúc phân cấp:** 3-5 cấp lồng nhau

3.2 Named Entities quan trọng

Các thực thể cần trích xuất:

- Tên môn học: IT001, SE104, IS217...
- Tên quyết định: QĐ 790, TB 1192, QĐ 1393...
- Số tín chỉ: 3 TC, 4 TC, 125 TC...
- Học kỳ/Khóa: Khóa 15, Khóa 2020...
- Chuẩn đầu ra: LO1, LO2, ..., LO10
- Ngày ban hành: 28/9/2022, 20/12/2022...

3.3 Cấu trúc phân cấp

Cấu trúc phân cấp rõ ràng giúp chunking thông minh:

- Quy định: CHƯƠNG I > Điều 1 > Khoản 1 > Mục a, b, c
- Chương trình: Phần > Mục > Tiểu mục > Bảng

3.4 Mối quan hệ ngữ nghĩa

- **Môn tiên quyết:** IT001 $\rightarrow$ IT002 $\rightarrow$ IT003
- **Khối kiến thức:** Phân loại theo cơ sở nhóm ngành, cơ sở ngành, chuyên ngành
- **Quy định thay thế:** QĐ 1393 (2023) thay thế một phần QĐ 790 (2022)
- **Chuẩn đầu ra - Môn học:** LO1 (Kiến thức nền tảng) $\rightarrow$ MA006, MA003, MA004

Chương 4 Vấn đề cần xử lý

4.1 Vấn đề với PDF (regulation)

4.1.1 Letterhead và Footer noise

Mỗi trang có header/footer lặp lại: logo, tên trường, số trang, chữ ký. Gây nhiễu cho embedding và retrieval.

4.1.2 Bảng biểu phức tạp

Bảng có merged cells, nhiều cột, nhiều cấp header. Parse dễ bị sai cấu trúc, thiếu thông tin, markdown table malformed.

4.1.3 Multi-column Layout

Văn bản 2 cột dễ bị parse sai thứ tự (đọc ngang thay vì dọc từng cột).

4.1.4 Chất lượng scan

Một số PDF là scan từ giấy, OCR có thể nhận dạng sai ký tự (ví dụ: ”Điều” thành ”Đjều”).

4.2 Vấn đề với Markdown (curriculum)

4.2.1 Navigation và Menu duplicates

Mỗi file MD crawl từ web có navigation menu, header/footer website lặp lại nhiều lần, không cần thiết.

4.2.2 Malformed Markdown

Bảng thiếu dấu |, số cột không đều. Heading không có space (##Heading thay vì ## Heading). List item không đúng indent.

4.2.3 Inconsistent Structure

Mỗi khóa/ngành có format hơi khác nhau:

- Khóa 2020: ”1.1. Mục tiêu đào tạo”

- Khóa 2022: ”1.1 Mục tiêu đào tạo”
- Khóa 2024: ”Mục tiêu đào tạo”

4.3 Vấn đề chung

4.3.1 Chunking Complexity

Document dài (100+ trang). Chunk cố định có thể cắt giữa ”Điều 5” và ”Khoản 2”, mất ngữ cảnh. Cần chunking thông minh theo semantic boundaries.

4.3.2 Semantic Overlap

Quy định mới thay thế/cập nhật quy định cũ, có thể có thông tin trùng lặp hoặc mâu thuẫn. Agent cần biết quy định nào mới nhất.

4.3.3 Low-quality Content

Một số page chỉ có header/footer, không có nội dung. File parse lỗi hoàn toàn. Nội dung quá ngắn (dưới 50 từ). Cần filter để reject.

Chương 5 Kết luận

Báo cáo này mô tả cấu trúc, đặc trưng và các vấn đề cần xử lý của dữ liệu cho hệ thống UIT AI Assistant. Dữ liệu bao gồm 27 file PDF quy định và 100+ file Markdown chương trình đào tạo, với các đặc trưng quan trọng như named entities, cấu trúc phân cấp và mối quan hệ ngữ nghĩa. Các vấn đề chính cần giải quyết là letterhead noise, bảng biểu phức tạp, malformed markdown và chunking complexity.