**Cách thu thập dataset và huấn luyện model cho chatbot tư vấn tuyển sinh**

**1. Thu thập dữ liệu đầu vào:**

* **Scraping website tuyển sinh:** Thu thập thông tin từ website chính thức của ĐH Cần Thơ (ngành học, điểm chuẩn, học phí, chính sách)
* **Phân tích fanpage/diễn đàn:** Lưu lại các câu hỏi-đáp phổ biến từ fanpage, nhóm sinh viên
* **Phỏng vấn cán bộ tuyển sinh:** Ghi lại những câu hỏi thường gặp và cách trả lời chuẩn
* **Google Form khảo sát:** Tạo form cho học sinh, phụ huynh để thu thập câu hỏi thực tế
* **Dữ liệu lịch sử chat:** Xin phép trường cung cấp dữ liệu chat/email tư vấn nếu có

**2. Xử lý và chuẩn bị dataset:**

* **Cấu trúc dữ liệu:** Tổ chức thành cặp <câu hỏi, câu trả lời>
* **Làm sạch dữ liệu:** Loại bỏ duplicate, sửa lỗi chính tả, chuẩn hóa dấu câu
* **Phân loại theo intent:** Chia thành các nhóm như:
* Thông tin ngành/khoa
* Điểm chuẩn/tuyển sinh
* Học phí/học bổng
* Cơ sở vật chất/ký túc xá
* Thủ tục hồ sơ
* **Tăng cường dữ liệu:** Viết lại câu hỏi với nhiều cách khác nhau (paraphrasing)
* **Gán nhãn entities:** Trích xuất entities quan trọng (tên ngành, khoa, mức điểm...)

**3. Chia tập dữ liệu:**

* Training (70%), Validation (15%), Testing (15%)

**4. Lựa chọn và huấn luyện model:**

* **Mô hình NLU:**
* Sử dụng mô hình có sẵn như BERT-Vietnamese, PhoBERT, hoặc RASA NLU
* Fine-tune với dữ liệu tuyển sinh của trường
* **Phương pháp huấn luyện:**
* Intent classification: Phân loại câu hỏi thuộc chủ đề nào
* Entity recognition: Nhận diện thông tin quan trọng trong câu hỏi
* Response generation: Tạo câu trả lời phù hợp
* **Hyperparameter tuning:** Điều chỉnh learning rate, batch size, epochs

**5. Đánh giá model:**

* Accuracy, Precision, Recall cho việc phân loại intent
* F1-score cho entity recognition
* BLEU score hoặc đánh giá thủ công cho phần sinh câu trả lời
* A/B testing với người dùng thật

**6. Cải thiện liên tục:**

* Thêm dữ liệu mới dựa trên các trường hợp model trả lời sai
* Thu thập feedback từ người dùng thực tế
* Cập nhật nội dung hàng năm theo quy chế tuyển sinh mới