**DỰ ÁN CUỐI KỲ**

Nhập môn Xử lý Ngôn ngữ Tự nhiên – 2024

**Bài 1 (5đ):**

Tìm hiểu và trình bày về các biểu diễn token theo phương pháp BPE (Byte-Pair Encoding). Cho các ví dụ về các mô hình có sử dụng tokenizer theo BPE.

So sánh 2 mô hình trên một bài toán cụ thể nào đó có sử dụng BPE và không sử dụng BPE. Lưu ý 2 mô hình này phải được train và test trên cùng một bộ dữ liệu.

**Bài 2 (5đ)**: Chọn 1 trong 2 bài sau:

Bài 2a:

Tìm hiểu và xây dựng bài toán phát hiện lỗi và sửa lỗi chính tả tiếng Việt theo tiếp cận học sâu. Các nhiệm vụ cần thực hiện:

1. Chọn mô hình: mô hình chỉ có Encoder; hoặc mô hình sinh Encoder-Decoder; hoặc mô hình chỉ sử dụng Decoder
2. Xây dựng, sinh dữ liệu huấn luyện dựa trên các loại lỗi chính tả tiếng Việt thường gặp.
3. Huấn luyện mô hình. Có thể sử dụng Pre-trained model và thực hiện fine-tune hoặc tự train từ đầu.
4. Hãy đánh giá độ chính xác của mô hình

Bài 2b: Hãy xây dựng mô hình để tạo một trợ lý ảo (agent) làm nhiệm vụ chăm sóc khách hàng trong một lĩnh vực nào đó tự bạn chọn, ví dụ:

* Agent của một siêu thị cho cho giới thiệu các mặt hàng và trả lời câu hỏi của khách hàng
* Agent chăm sóc khách hàng cho khách sạn
* Agent chăm sóc khách hàng cho công ty du lịch
* … (tự bạn chọn lĩnh vực)

Các nhiệm vụ cần làm:

* Thu thập và xây dựng dữ liệu huấn luyện
* Lựa chọn mô hình huấn luyện
* Đánh giá độ chính xác của mô hình

**Nộp bài: Đối với cả bài 1 và bài 2**

* File Word chứa trình bày về lý thuyết
* File codes