**Dự án Cuối kỳ**

Môn: Nhập môn Xử lý Ngôn ngữ tự nhiên

**Bài 1 (5đ):**

1. Hãy trình bày cơ chế Attention trong các mô hình Sequence-to-Sequence bao gồm:
   * Mô hình dựa trên LSTM
   * Mô hình Transformer
   * Mô hình GPT (Generative Pretrain Transformer)

Phân tích sự khác nhau về Attention giữa các kiến trúc này.

1. Xây dựng mô hình Lstm-to-Lstm có attention và không có attention cho bài toán Dịch máy từ ngôn ngữ tiếng Việt sang tiếng Anh:

- Hãy so sánh kết quả giữa 2 mô hình này.

- Chỉ ra attention nó được thể hiện như thế nào qua 1 ví dụ đưa vào.

**Bài 2 (5đ):**

1. Chọn 1 bài toán cụ thể thuộc loại trích chọn thông tin (Information Extraction), chọn bài toán có ý nghĩa trong thực tế. Và ưu tiện chọn loại bài toán mới.

*Input:* 1 đoạn văn bản (có thể là 1 câu)

*Output:* thông tin cần trích chọn và format

Ví dụ cho bài toán NER:

Input: “ông Trần Văn Nam đang đi chơi tại Hà Nội”

Output: “<person> Trần Văn Nam <place> Hà Nội”

Hãy tự xây dựng dữ liệu để huấn luyện và dữ liệu Test.

Gợi ý: có thể dùng các mô hình ngôn ngữ lớn để xây dựng dữ liệu.

1. Xây dựng mô hình Transformer (có thể sử dụng Pretrained model) để huấn luyện và đánh giá bài toán trên.

Lưu ý: đánh gía trên bộ Test với các độ đo Precison, Recall, F1-score

Hướng dẫn nộp bài:

* File pdf chứa phần lý thuyết
* Các file code python để chạy huấn luyện và test
* File data chứa dữ liệu

**Lưu ý:**

* Tất cả các các thành viên trong nhóm đều nộp bài lên Elearning;
* Tất cả các thành viên đều cùng làm bài, cùng hiểu tất cả các bài. Tránh trường hợp phân chia mỗi người chỉ hiểu phần mình làm.