TIỂU LUẬN
XỬ LÝ NGÔN NGỮ TỰ NHIÊN

QUY ĐỊNH

Bài làm cần nộp lên hệ thống eLearning của khoa từ ngày **10/6/2025** đến hết ngày **30/6/2025**. **Bài nộp muộn sẽ không được chấp nhận.**

Thời gian thuyết trình bắt đầu từ ngày **23/6/2025**. Tất cả các thành viên trong nhóm đều phải tham gia thuyết trình.

Bài làm cần tuân thủ các quy định sau:

- Đây là bài làm nhóm. Mỗi nhóm không được quá 3 thành viên.
- Các trường hợp đạo văn sẽ bị 0 điểm và xử lý theo quy định của khoa.
- Bài nộp phải được trình bày rõ ràng. Nếu vi phạm định dạng hoặc trình bày cẩu thả, bài làm có thể bị trừ từ 10% đến 50% tổng điểm.
- Mọi thắc mắc cần giải đáp liên hệ giảng viên phụ trách lớp.

Nộp bài:

- Học viên cần nộp:
    - Bài báo cáo dưới dạng file Word **và** PDF **(**.doc/.docx và .pdf**).**
    - Mã nguồn Python.
    - Slide báo cáo.
- Tất cả các file được nén và đặt tên file nén theo **Mã số học viên** của nhóm, ví dụ: **52200000_52200001.docx**.
- Bài làm phải tuân theo quy định của khoa**,** không tính trang bìa, danh mục tài liệu tham khảo và mục lục.
- Thời hạn nộp bài:
    - Lần 1: Nộp báo cáo sơ bộ (docx, pdf) và file trình bày (ppt). Hạn nộp bài trước 17h ngày 22/6/2025.
    - Lần 2: Nộp báo cáo hoàn chỉnh (docx, pdf) và mã nguồn các chương trình. Hạn cuối 23h59’ ngày 30/6/2025.

NỘI DUNG BÀI LÀM

1. Tìm hiểu và thực nghiệm cách biểu diễn Doc2vec sử dụng Deep Learning:
   a. Có sử dụng pretrained models.
   b. Không sử dụng pretrained models.

2. Tìm hiểu ít nhất 02 thuật toán Reinforcement Learning sử dụng trong các mô hình ngôn ngữ lớn.
   - Trình bày công thức, giải thuật, giải thích ý nghĩa, so sánh các thuật toán với nhau.
   - Có code minh họa cho mỗi thuật toán.
   - Dữ liệu huấn luyện cần thiết.

3. Giải quyết 1 trong 2 bài toán sau:
   a. Xây dựng mô hình dịch các câu thông dụng từ tiếng Anh sang tiếng Việt.
   b. Mô hình hỏi đáp trong một lĩnh vực nào đó, ví dụ tư vấn y tế, tư vấn giáo dục, tư vấn tình cảm, …

   Tương ứng với mỗi bài toán cần thực hiện các yêu cầu sau:
   - Mô hình Encoder-Decoder Transformer.
   - Mô hình GPT (Generative Pretrained Transformer)
   - Xây dựng từ pretrained models và train từ đầu.
   - Trình bày phần tokenizer, thuật toán BPE.
   - Trình bày và thực nghiệm các độ đo để đánh giá mô hình.
