BÀI THỰC HÀNH 1: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

Hướng dẫn nộp bài:

Nội dung nộp sẽ bao gồm:

- 01 file báo cáo (theo mẫu).
- 01 file zip chứa dữ liệu đã thu thập và trước khi tiền xử lý.
- 01 file zip chứa dữ liệu đã thu thập và sau khi tiền xử lý.

Đặt tên là **MSSV_BaiThucHanh1.zip**, nhóm cử 1 bạn thành viên đại diện nộp bài lên course.

Các bài toán đề xuất:

1. Đọc hiểu tự động trên bài báo điện tử về sức khoẻ trong tiếng Việt (Paper: https://arxiv.org/abs/2006.11138).
2. Nhận diện bình luận mang tính đóng góp và bình luận mang tính toxic trên mạng xã hội (Paper: https://arxiv.org/pdf/2103.10069.pdf).
3. Nhận diện cảm xúc của câu bình luận trên các nền tảng mạng xã hội (Paper: https://arxiv.org/pdf/1911.09339.pdf).
4. Hỏi đáp tự động dựa trên hình ảnh trong tiếng Việt (Paper: https://arxiv.org/abs/2305.04183).
5. Nhận diện bình luận độc hại trên mạng xã hội (Paper: https://arxiv.org/abs/2103.11528).
6. Nhận diện bình luận phàn nàn trên mạng xã hội (Paper: https://arxiv.org/pdf/2104.11969.pdf)
7. Nhận diện tin nhiễu trong các bài đăng tin quảng cáo bất động sản (Paper: https://aclanthology.org/2022.paclic-1.71.pdf).

Yêu cầu:

- Các bạn sinh viên chọn nhóm tối thiểu 2 bạn tối đa 5 bạn cho Bài Thực hành 1 và Bài Thực hành 2.
- Nếu có nhiều hơn một nhóm chọn cùng 1 bài toán thì cần đảm bảo dữ liệu thu thập của hai nhóm đôi một không có quá 10% samples trùng lặp.
- Hoàn thành các bài tập được giao.

Lưu ý: Bài Thực hành 1 tập trung thu thập dữ liệu chuẩn bị cho quy trình gán nhãn, chưa đề cập đến việc gán nhãn cho bộ dữ liệu. Mọi chi tiết thắc mắc mạnh dạn liên hệ qua email 19520178@gm.uit.edu.vn hoặc https://www.facebook.com/hieunghia.nhn để được giải đáp.

Bài 1: Chọn 01 bài toán và thu thập 500 samples mẫu cho bài toán tương ứng.

Bài 2: Tiến hành tiền xử lý dữ liệu (làm sạch dữ liệu, fill missing values, tổ chức cấu trúc lại dữ liệu).