THỰC HÀNH 4: PHÂN LỚP VĂN BẢN

Hướng dẫn nộp bài:

Nội dung nộp sẽ bao gồm:

- 01 file báo cáo.
- 01 file zip chứa toàn bộ source code.

Đặt tên là **MSSV_BaiThucHanh4.zip** và nộp lên course.

Bài toán: Nhận dạng cảm xúc từ câu phản hồi bằng văn bản của người dùng.

Input: câu bình luận của người dùng.

Output: Nhãn cảm xúc, gồm 1 trong 3 nhãn: tích cực, tiêu cực và trung tính.

Bộ dữ liệu: **UIT-VSFC**

Paper: **UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis** (KSE 2018).

Link: https://drive.google.com/drive/folders/1xclbjHHK58zk2X6iqbvMPS2rcy9y9E0X?usp=drive_open.

Bài 1: Mã hoá văn bản bằng kỹ thuật Counting. Huấn luyện phương pháp Naive Bayes, SVM và Softmax Regression. Sử dụng các độ đo accuracy, precision, recall và F1 để đánh giá mô hình huấn luyện được.

Bài 2: Mã hoá văn bản bằng kỹ thuật TF-IDF. Huấn luyện phương pháp Naive Bayes, SVM và Softmax Regression. Sử dụng các độ đo accuracy, precision, recall và F1 để đánh giá mô hình huấn luyện được.

Bài 3: Sử dụng VnCoreNLP thực hiện tách từ. Mã hoá văn bản bằng kỹ thuật TF-IDF. Huấn luyện phương pháp Naive Bayes, SVM và Softmax Regression. Sử dụng các độ đo accuracy, precision, recall và F1 để đánh giá mô hình huấn luyện được. Nhận xét ảnh hưởng của việc tách từ lên kết quả của mô hình SVM.

Bài 4: Tìm hiểu thêm một số kỹ thuật tiền xử lý, làm sạch dữ liệu để cải thiện chất lượng dữ liệu và hiệu suất của các mô hình máy học. (Có thể xem tham khảo **source code**)