# BÀI TẬP LỚN: PHÂN TÍCH PHẢN HỒI SINH VIÊN BẰNG NLP VỚI BỘ DỮ LIỆU UIT-VSFC

Môn học: Lập trình Python với khoa học dữ liệu

Hạn nộp bài: 23:59 ngày 21/6/2025.

Hình thức nộp:

* Tệp .ipynb hoặc Colab
* Báo cáo bản cứng kèm file báo cáo bản mềm định dạng .docx hoặc .pdf
* Thuyết trình kết quả (dự kiến trong học buổi sáng 21/6/2025)

## Mục tiêu:

Phân tích dữ liệu phản hồi của sinh viên sử dụng NLP (xử lý ngôn ngữ tự nhiên), áp dụng các kỹ thuật khai phá dữ liệu, phân loại cảm xúc và chủ đề trong văn bản, đánh giá chất lượng giảng dạy dựa trên phản hồi thu thập được.

## Dữ liệu:

* Tên: UIT-VSFC (Vietnamese Students' Feedback Corpus)
* Ý nghĩa: Phản hồi của sinh viên đối với một chương trình đào tạo
* Nguồn: <https://huggingface.co/datasets/uitnlp/vietnamese_students_feedback>
* Tham chiếu trong tài liệu báo cáo:

@InProceedings{8573337,

author={Nguyen, Kiet Van and Nguyen, Vu Duc and Nguyen, Phu X. V. and Truong, Tham T. H. and Nguyen, Ngan Luu-Thuy},

booktitle={2018 10th International Conference on Knowledge and Systems Engineering (KSE)},

title={UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis},

year={2018},

pages={19-24},

doi={10.1109/KSE.2018.8573337}

}

Dữ liệu gồm 16.175 phản hồi của sinh viên, bao gồm các trường:

* sentence: Câu phản hồi
* sentiment: Cảm xúc khi phản hồi (0: tiêu cực, 1: trung lập, 2: tích cực)
* topic: Chủ đề phản hồi (0: giảng viên, 1: chương trình, 2: cơ sở vật chất, 3: khác)

Dữ liệu đã được chia thành 3 nhóm: Train (11426 bản ghi), Validate (1583 bản ghi) và Test (3166 bản ghi). Với mỗi nhóm sẽ gồm 3 tệp với tên <tên nhóm>-sentences, <tên nhóm>-sentiments và <tên nhóm>-test (để tránh lỗi dữ liệu do câu phản hồi có thể chứa dấu “,”).

## Nội dung và yêu cầu bài làm:

1. Tiền xử lý dữ liệu

* Làm sạch văn bản: Loại bỏ ký tự đặc biệt, chuẩn hóa chữ hoa/thường.
* Tokenization: Sử dụng thư viện phù hợp cho tiếng Việt để tách từ.
* Loại bỏ stopwords nếu cần thiết.

1. Khám phá dữ liệu (EDA)

* Thống kê số lượng phản hồi theo từng nhãn cảm xúc và chủ đề.
* Trực quan hóa phân bố cảm xúc và chủ đề bằng biểu đồ phù hợp.
* Phân tích tần suất các từ theo nhãn cảm xúc.
* Xác định cụm từ đặc trưng cho phản hồi tích cực/trung tính/tiêu cực.

1. Phân loại cảm xúc

Với các chia bộ dữ liệu được cung cấp, áp dụng một số mô hình sau để tiến hành phân loại, dự đoán cảm xúc:

* Mô hình truyền thống: Naive Bayes, SVM, TF-IDF.
* Mô hình học sâu: LSTM, CNN.
* Mô hình transformer: PhoBERT.
* Đánh giá hiệu suất các mô hình trên cơ sở các chỉ số: Accuracy, Precision, Recall, F1-score. Lựa chọn 2-3 mô hình tốt nhất.

1. Phân loại chủ đề

* Làm tương tự như trên nhưng với áp dụng cho chủ đề phản hồi.

1. Kiểm tra tính ổn định của các mô hình được lựa chọn ở câu 4-5 bằng cách chia lại bộ dữ liệu (Train, Validate, Test) theo cách (ngẫu nhiên) khác.
2. Xây dựng mô hình để phát hiện các phản hồi có mức độ quan trọng, cần sự quan tâm, xử lý (với phản hồi tiêu cực) hoặc truyền thông, nhân rộng (với phản hồi tích cực) từ phía nhà trường.

Gợi ý: Huấn luyện binary classifier với 2 đích là Phản hồi thường vs Phản hồi quan trọng. Có thể xây dựng bộ từ điển thể hiện cảm xúc mạnh (vd: “bức xúc”, “không thể chấp nhận”, “tuyệt vời”, “rất hài lòng”…).

## Kết quả mong đợi:

* Một Notebook hoặc Colab có đầy đủ mã nguồn, biểu đồ, giải thích rõ ràng
* Một báo cáo trình bày các phát hiện, trực quan hóa, bảng biểu và kết luận
* Một buổi thuyết trình ngắn (15-20’) trình bày và thảo luận kết quả.

## Hình thức nộp bài:

* File nén .zip gồm:  
   • Tệp .ipynb hoặc link Colab  
   • Báo cáo .pdf hoặc .docx
* Nộp qua hệ thống LMS/lớp học hoặc email giảng viên trước deadline.

## Tiêu chí đánh giá (100 điểm)

|  |  |
| --- | --- |
| **Tiêu chí** | **Trọng số** |
| Tiền xử lý dữ liệu | 15 |
| Phân tích dữ liệu (EDA) | 15 |
| Mô hình phân loại cảm xúc | 30 |
| Mô hình phân loại chủ đề | 20 |
| Phát hiện phản hồi quan trọng | 10 |
| Trình bày báo cáo & code | 10 |