# Lab 3 - Phân loại thư rác

## 1 Tập dữ liệu Enron-Spam

### 1.1 Giới thiệu

- Bộ dữ liệu Enron-Spam là một nguồn tài liệu tuyệt vời được thu thập bởi V. Metsis, I. Androutsopoulos và G. Paliouras và được mô tả trong ấn phẩm của họ "Spam Filtering with Naive Bayes - Which Naive Bayes?". Bộ dữ liệu chứa tổng cộng 17.171 thư rác và 16.545 thư không phải thư rác ("ham") (tổng cộng 33.716 thư điện tử).
- Mỗi thư (1 dòng) trong tập dữ liệu có đặc điểm như sau:
  + Subject: tên tiêu đề của thư.
  + Message: Nội dung của email. Có thể chứa chuỗi rỗng nếu tin nhắn chỉ có dòng tiêu đề và không có nội dung. Trong trường hợp chuyển tiếp email hoặc trả lời, điều này cũng chứa tin nhắn gốc với dòng tiêu đề, "từ:", "đến:", v.v.
  + Spam/Ham: Có giá trị "spam" hoặc "ham". Nhãn của thư được phân loại có là tin nhắn spam hay không.

### 1.2 Tải tập dữ liệu

- Tải từ moodle.
- Tải từ link drive: [link]
- Thư mục dữ liệu bao gồm 2 file: train.csv và val.csv
  - train.csv: Gồm những dữ liệu dùng để huấn luyện mô hình.
  - val.csv: Gồm những dữ liệu để đánh giá mô hình sau khi train.

### 1.3 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Tải xuống và đọc được toàn bộ tập dữ liệu Enron-Spam.
- Đọc dữ liệu từ file và in ra **5** dòng đầu tiên của tập dữ liệu.

## 2 Tiền xử lý dữ liệu (3 điểm)

### 2.1 Về dữ liệu

- Nội dung dữ liệu:
  - Đầu vào gồm có 2 đặc trưng: Subject, Message.
  - Đầu ra là một nhãn của email: Spam hoặc Ham.

### 2.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- (Optional) Làm sạch dữ liệu nếu có thể: kiểm tra những dòng bị lặp, biến đổi/sắp xếp thứ tự các dòng dữ liệu lại,...
- Các nhóm được phép sử dụng thêm một số phương pháp khác Để gia tăng hiệu quả của mô hình nhưng nhóm cần cho biết phương pháp mình áp dụng là gì và cho biết mức độ cải thiện cụ thể gia tăng bao nhiêu.

# 3 Các yêu cầu về mô hình (5 điểm)

## 3.1 Mô hình

Ở lớp lý thuyết sinh viên đã được học về Maximum Likelihood estimation (MLE) và Maximum A Posteriori estimation (MAP). Vì vậy, ở lab này, nhóm được yêu cầu sử dụng các mô hình thống kê (probabilistic models) để phân loại thư rác (tham khảo naive bayes classifier)

## 3.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Trình bày cấu trúc và cách thiết kế mô hình mình chọn một cách cụ thể, chi tiết từng bước tính toán từ đầu vào cho đến đầu ra.
- Điểm cộng (1 điểm): nhóm tự lập trình lại naive bayes classifer dựa vào kiến thức về MLE và MAP đã được học.
- Trong mã nguồn, nếu nhóm sử dụng các tham số đặc biệt nào đó thì cần tìm hiểu và giải thích lý do tại sao chọn.
- Sau khi huấn luyện, cho biết độ chính xác của mô hình đối với toàn bộ tập dữ liệu (bao gồm cả tập trainset và valset). Có thể sử dụng nhiều cách để đánh giá và cần giải thích những đánh giá ấy có ý nghĩa gì.

# 4 Thử nghiệm thực tế

## 4.1 Mục đích

Mặc dù ta đã có tập dữ liệu val set để đánh giá mô hình, nhưng ta vẫn muốn thử khả năng của mô hình bằng cách viết email trực tiếp.

## 4.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Chức năng 1: Viết một code block cho phép người dùng nhập vào một email bất kì (gồm tiêu đề và nội dung). Sau khi áp dụng các phương pháp tiền xử lý giống như đã làm với tập dữ liệu, chương trình chạy mô hình và trả ra kết quả dự đoán cho email vừa nhập.
- Chức năng 2: Viết một code block cho phép đọc một file "csv" (Comma seperated value) bất kì có cấu trúc như file val.csv (Bao gồm thông tin tiêu đề, nội dung và nhãn cho từng email) và thực hiện đánh giá kết quả dự đoán như đã thực hiện với tập val.csv.

# 5 Các yêu cầu khác

- Ngôn ngữ sử dụng bắt buộc là Python, không được phép sử dụng ngôn ngữ khác. (nên sử dụng Jupiter Notebook).
- Giới hạn thư viện: nhóm chỉ được sử dụng các thư viện cho các tác vụ nằm ngoài việc huấn luyện mô hình (ví dụ: pandas, numpy,...) và không được sử dụng các thư việc cho tác vụ này (ví dụ: sklearn,...)

- Các nhóm cần kiểm tra mã nguồn trước khi nộp. Nếu mã nguồn không chạy được mà không phải do nguyên nhân khách quan (thiếu thư viện, lỗi do thư viện gây ra, sử dụng thư viện sai phiên bản,...) thì sẽ bị 0 điểm đồ án.
- Bài nộp phải gồm có 2 phần:
  + Report: Chứa các file báo cáo. **Giới hạn tối đa 10 trang**.
  + Source: Chứa các file mã nguồn.
- Trong các file nộp, nhóm cần ghi rõ thông tin về các thành viên gồm họ tên và MSSV. Riêng đối với mã nguồn, nhóm có thể ghi thông tin trên dưới dạng comment trong code của nhóm.
- Bài nộp sẽ được đặt trong thư mục có tên `MSSV01[_MSSV02[_MSSV03[...]]]` và được nén lại bằng định dạng ZIP với format `[group_number]`.zip . Ví dụ đặt tên nhóm có 1 nhóm là `MSSV01`, nhóm có 2 nhóm là `MSSV01_MSSV02`.
- Nghiêm cấm các hành vi gian lận, không trung thực trong học tập như sao chép bài làm giữa các nhóm với nhau, sao chép bài làm của các nhóm khóa trước hoặc các nhóm lớp khác trường khác, nhờ người làm hộ. Nếu phát hiện các hành vi trên thì cả nhóm sẽ bị 0 điểm và xử lý theo quy định của Khoa và Trường.