Báo cáo thực hành buổi 3  
Họ và tên: Lương Thành An Lớp: 64TTNT1 MSV:2251262567

Link github: https://github.com/Annlt272/TH3.git

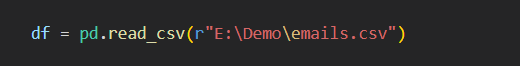
Chuẩn bị dữ liệu.

1. Link data: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset
2. Dữ liệu (jena\_climate\_2009\_2016) gồm:

- Nhiệt độ, độ ẩm, gió, áp suất mỗi 10 phút

- Ghi chép liên tục 4 năm

II. Mô tả công việc đã làm.



Đoạn code này thực hiện việc đọc dữ liệu từ email.csv chứa thông tin hơn 500.000 email từ 150 nhân viên của Tập đoàn Enron. Đây là một trong những bộ dữ liệu email công khai lớn nhất và đã được sử dụng rộng rãi trong nghiên cứu về xử lý ngôn ngữ tự nhiên, phân tích mạng xã hội và học máy.

A computer screen with text

AI-generated content may be incorrect.

Hàm trích ngày gửi từ chuỗi message dùng regex: Tìm đoạn giữa Date và From, dùng re.DOTALL để khớp qua nhiều dòng nếu cần. Trả về chuỗi thời gian nếu tìm thấy.

A screen shot of a computer code

AI-generated content may be incorrect.

Áp dụng hàm extract\_date để tạo cột date\_str chứa ngày gửi.

Chuyển chuỗi thời gian sang kiểu datetime. Nếu lỗi (format sai), gán giá trị là NaT.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Hàm loại bỏ múi giờ (timezone) nếu có, để chuẩn hóa thời gian.



Trích xuất đặc trưng thời gian: Tạo cột hour dạng số thực từ thời gian gửi email, ví dụ: 14h30 → 14.5.

Huấn luyện mô hình GMM:

A black background with white text

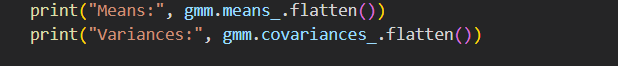
AI-generated content may be incorrect.

Chuyển cột hour thành mảng 2 chiều để đưa vào mô hình GMM.

A black background with white text

AI-generated content may be incorrect.

Khởi tạo GMM với 3 cụm. Huấn luyện và phân cụm các giờ gửi email. Gán nhãn cụm vào cột cluster.



In ra trung bình (mean) và phương sai (variance) của từng cụm.

A screen shot of a computer code

AI-generated content may be incorrect.

log\_likelihood: mức độ phù hợp của mô hình với dữ liệu. AIC, BIC: tiêu chí chọn mô hình tốt nhất (càng thấp càng tốt). In các chỉ số đánh giá mô hình.

A screenshot of a computer

AI-generated content may be incorrect.

Vẽ biểu đồ histogram giờ gửi email:

A computer screen shot of colorful text

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

Số lượng email không được gửi đều trong ngày. Có những khung giờ cao điểm và thấp điểm rất rõ rệt. Đường cong màu đỏ mô hình hoá mật độ gửi email qua GMM với 3 cụm:đỉnh 1 khoảng 2–5 giờ, đỉnh 2 khoảng 17–19 giờ chiều, đỉnh 3 khoảng 21–23 giờ tối.

III. Mô tả kết quả đạt được.

Trong quá trình phân tích dữ liệu từ tập **Enron Email Dataset**, mục tiêu chính là tìm hiểu hành vi gửi email theo thời gian trong ngày. Tập dữ liệu gồm hơn 500.000 email, trong đó trường message chứa thông tin tiêu đề, thời gian và nội dung. Chúng tôi đã trích xuất thời gian gửi email từ nội dung thô và chuyển đổi sang định dạng datetime, sau đó loại bỏ các giá trị không hợp lệ. Từ dữ liệu thời gian này, chúng tôi tính toán mốc **giờ trong ngày** (ví dụ: 13.5 tương ứng với 13 giờ 30 phút) và áp dụng mô hình **Gaussian Mixture Model (GMM)** để phân cụm. Kết quả cho thấy mô hình GMM với 3 cụm là phù hợp để mô tả phân phối thời gian gửi email. Ba cụm tương ứng với ba khoảng thời gian hoạt động rõ rệt trong ngày: cụm đầu tiên vào khoảng 2–5 giờ sáng, cụm thứ hai vào khoảng 17–19 giờ chiều, và cụm thứ ba vào khoảng 21–23 giờ tối. Biểu đồ trực quan thể hiện rõ các đỉnh trong phân phối, cho thấy lượng lớn email được gửi vào các khung giờ này. Trong khi đó, khoảng thời gian từ 9h đến 14h có mật độ gửi email thấp hơn hẳn, cho thấy một khoảng lặng trong ngày làm việc. Kết quả phân tích cung cấp cái nhìn sâu sắc về nhịp độ làm việc và thói quen gửi email của các nhân viên trong tổ chức Enron, đồng thời mở ra hướng nghiên cứu thêm như phân tích theo ngày trong tuần, theo người gửi hoặc nội dung email.