**THỰC HÀNH BUỔI 5**

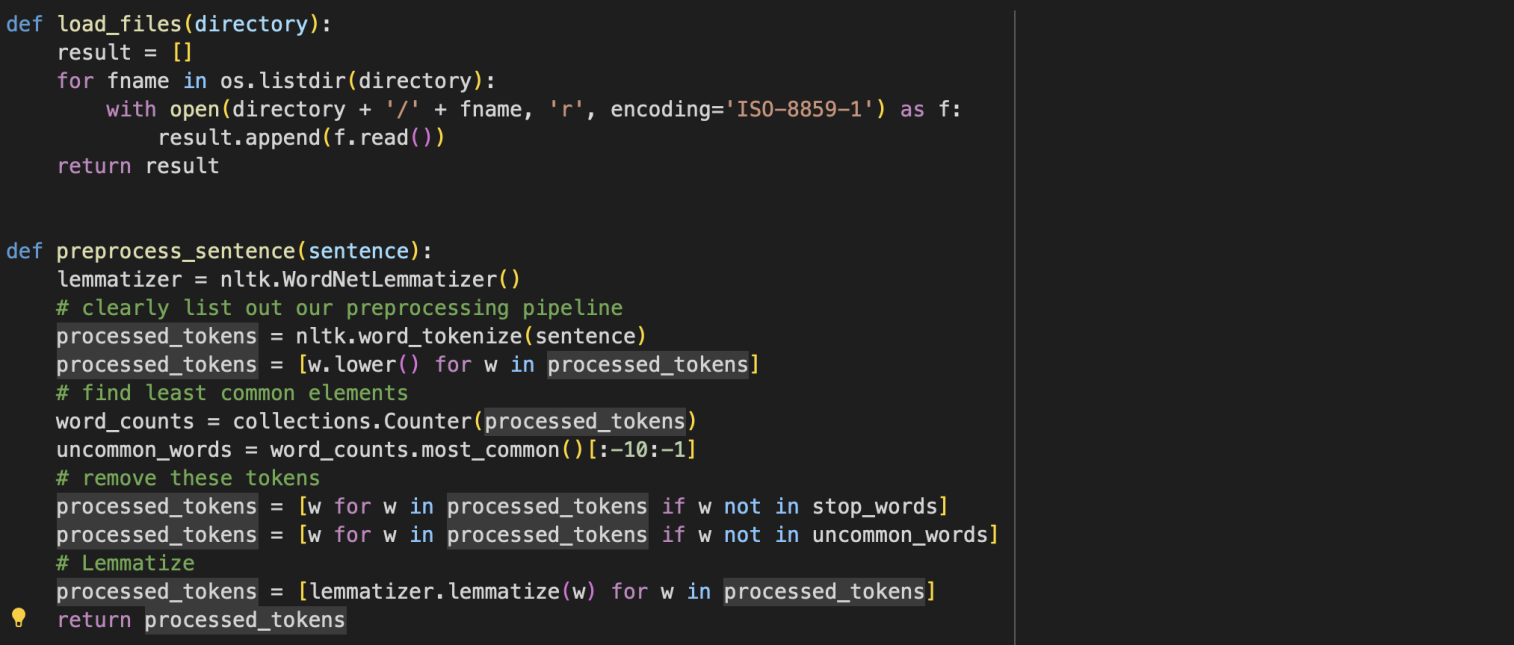
**TEXT CLASSIFICATION**

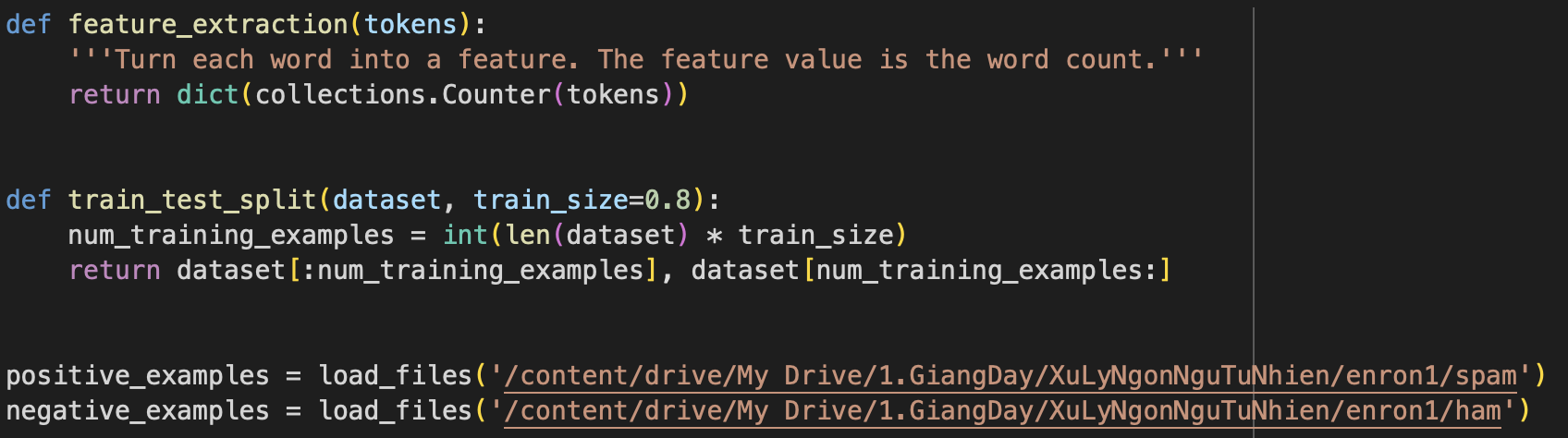
|  |
| --- |
| Họ tên và MSSV:  Lớp: Ngày học: |

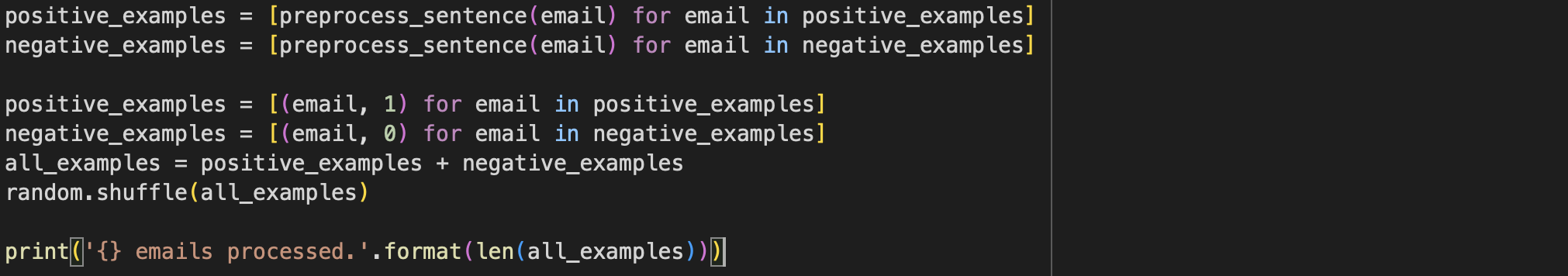
1. **Thực hành code**

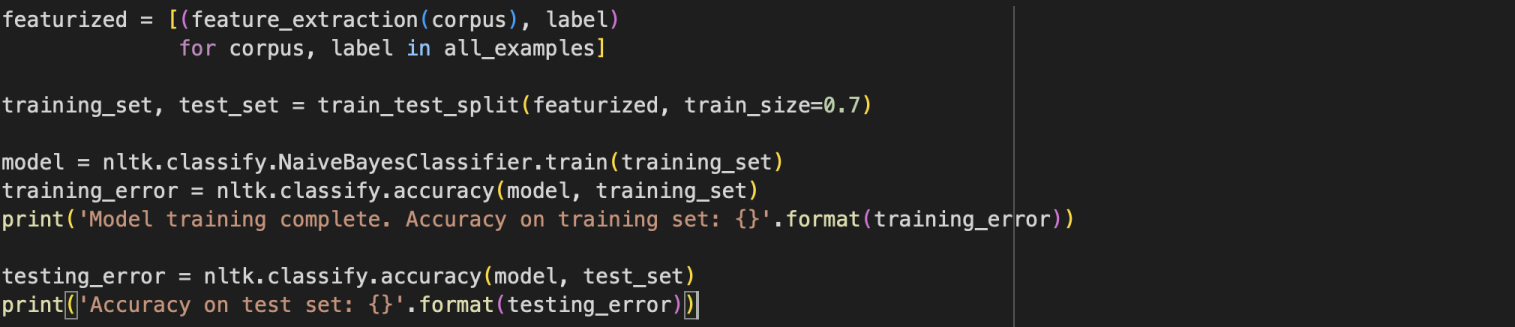
**Dataset:** https://www2.aueb.gr/users/ion/data/enron-spam/











1. **Bài tập**

Sau khi thực hiện các đoạn code trên, sinh viên trả lời các câu hỏi sau:

1. Đoạn code trên được dùng để làm gì, ý nghĩa của từng hàm và các bước thực hiện.
2. Mục đích của việc sử dụng WordNetLemmatizer trong tiền xử lý văn bản là gì?
3. Trong đoạn mã train\_test\_split, tham số train\_size=0.8 có ý nghĩa gì?
4. Tại sao chúng ta cần xáo trộn dữ liệu trước khi chia tập huấn luyện và kiểm tra?
5. Mô hình sử dụng thuật toán nào để phân loại email spam?
6. Nếu muốn cải thiện độ chính xác của mô hình, bạn có thể thực hiện những thay đổi nào?
7. Vì sao Naive Bayes lại phù hợp với bài toán phân loại văn bản?

Sinh viên nhập câu trả lời tại đây

|  |
| --- |
| **FILE PHẢI NỘP CHO BÀI THỰC HÀNH NÀY** |
| **Một file Word** chứa đáp án các yêu cầu |

*--- Hết ---*