TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**NHẬN DIỆN CẢM XÚC CHO VĂN BẢN TIẾNG VIỆT**

*Người hướng dẫn*: **PhD LÊ CUNG T ƯỞNG**

*Người thực hiện*: **LÊ HOÀNG LONG – 518H0035**

**TRẦN VĂN AN – 518H0127**

Lớp **: 18H50205**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**NHẬN DIỆN CẢM XÚC CHO VĂN BẢN TIẾNG VIỆT**

Người hướng dẫn: **PhD LÊ CUNG T ƯỞNG**

Người thực hiện: **LÊ HOÀNG LONG**

**TRẦN VĂN AN**

Lớp **: 18H50205**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Chung em xin gửi lời cảm ơn chân thành nhất đến với khoa nói chung cũng như Giảng viên Lê Cung Tưởng nói riêng, đã dành thời gian cũng như tâm huyết để dạy dỗ chúng em.

Đây là bài làm của chúng em. Trong quá trình làm khó tránh khỏi những sai xót. Chúng em rất mong nhận được lời góp ý từ mọi người để có thể ngày càng hoàn thiện thêm về mặt kiến.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của TS Nguyễn Văn A;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lê Hoàng Long*

*Trần Văn An*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Hoàn thành tốt task 1 và 2.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc72094155)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc72094156)

[TÓM TẮT iv](#_Toc72094157)

[MỤC LỤC 1](#_Toc72094158)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc72094159)

[CHƯƠNG 1 4](#_Toc72094160)

[GIỚI THIỆU 4](#_Toc72094161)

[1.1 Đề bài 4](#_Toc72094162)

[1.2 Tổng quan: 4](#_Toc72094163)

[CHƯƠNG 2 5](#_Toc72094164)

[CODE DEMO 5](#_Toc72094165)

[1.1 Task1: 5](#_Toc72094166)

[1.2 Task2: 8](#_Toc72094167)

[CHƯƠNG 3 15](#_Toc72094168)

[KẾT QUẢ 15](#_Toc72094169)

[3.1 Task1: 15](#_Toc72094170)

[3.2 Task2: 17](#_Toc72094171)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

CHƯƠNG 1

GIỚI THIỆU

1.1 Đề bài

Task 1. (40%) Thu thập dữ liệu 1. Chọn ba video Việt Nam trên YouTube có ít nhất 100 bình luận. Thu thập thông tin 100 nhận xét đầu tiên của mỗi video. Vui lòng tìm Ví dụ về thu thập dữ liệu.ipynb để biết ví dụ về thu thập thông tin. 2. Tự dán nhãn những bình luận này theo bảy lớp: Chán ghét (Chán ghét), Thích thú (thích thú), Giận dữ (chọc dữ), Ngạc nhiên (tức nhiên), Buồn bã (buồn bã), Sợ hãi (sợ hãi), Khác (khác).

Task 2. (30%) Nhận biết cảm xúc đối với văn bản trên mạng xã hội Việt Nam 1. Sử dụng bộ đào tạo và xác nhận của UIT-VSMEC (UIT-VSMEC.zip) để đào tạo mô hình Nhận biết cảm xúc của bạn (mô hình phân loại). Lưu ý: bạn có thể chọn bất kỳ kỹ thuật nào trong tiền xử lý, bất kỳ trình phân loại nào bạn muốn. 2. Báo cáo các chỉ số hiệu suất (Độ chính xác, điểm F1 ...) cho bộ kiểm tra trong bộ dữ liệu UIT- VSMEC.

Task 3. (30%) Áp dụng mô hình đã đào tạo trong Nhiệm vụ 2 cho ba tập dữ liệu trong Nhiệm vụ 1. Sau đó, báo cáo các chỉ số đo lường hiệu suất (Độ chính xác, điểm F1 ...) cho các tập dữ liệu này.

1.2 Tổng quan:

- Nhận dạng cảm xúc là một trường hợp đặc biệt của phân tích cảm xúc.

- Nhận dạng cảm xúc đóng vai trò quan trọng trong việc đo lường giá trị thương hiệu của một sản phẩm và có nhiều vai trong quan trọng trong nhiều lĩnh vực khác.

- Paul Ekman đề xuất 6 cảm xúc cơ bản của con người bao gồm: Disgust (chán ghét), Enjoyment (thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi).

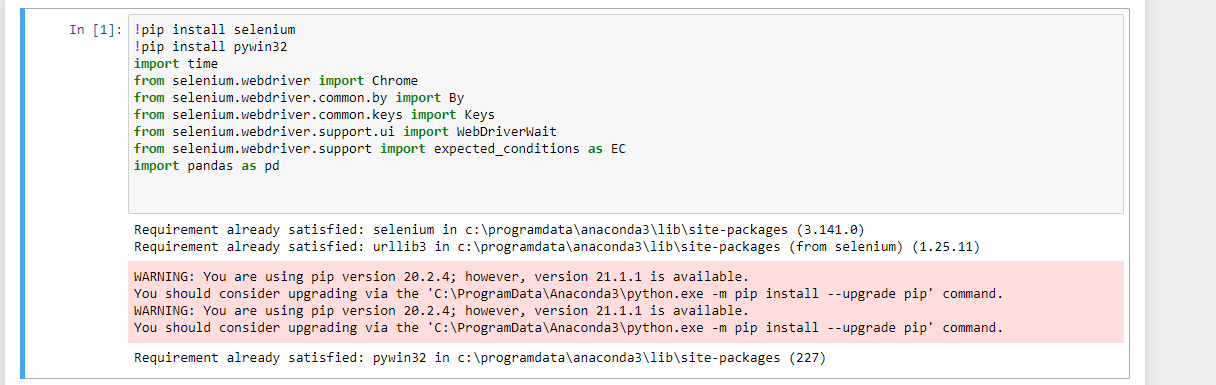
- Trong những năm gần đây, nhận dạng cảm xúc trong văn bản đã trở nên nhiều hơn phổ biến do các ứng dụng tiềm năng rộng lớn của nó trong tiếp thị, bảo mật, tâm lý học, tương tác giữa con người với máy tính, trí tuệ nhân tạo, v.v.

CHƯƠNG 2

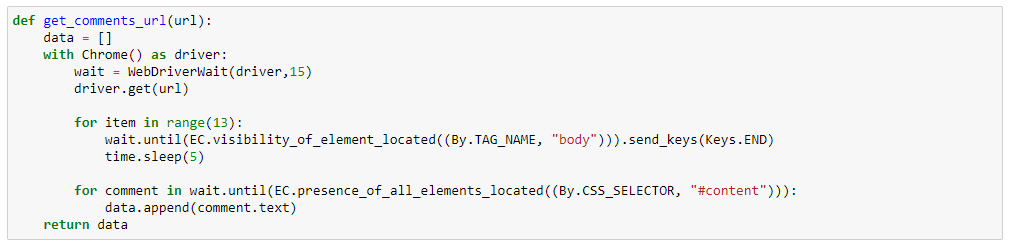
CODE DEMO

* 1. Task1:

Đầu tiên ta cần khai báo một số thư viện cần để crawl dữ liệu



Tiếp theo ta sẽ viết hàm để lấy comment từ url:



Sau đó ta sẽ gọi hàm lấy comment từ url và ở đây chúng em sẽ gọi hàm và truyền url và với hàm tương ứng với 3 tập data



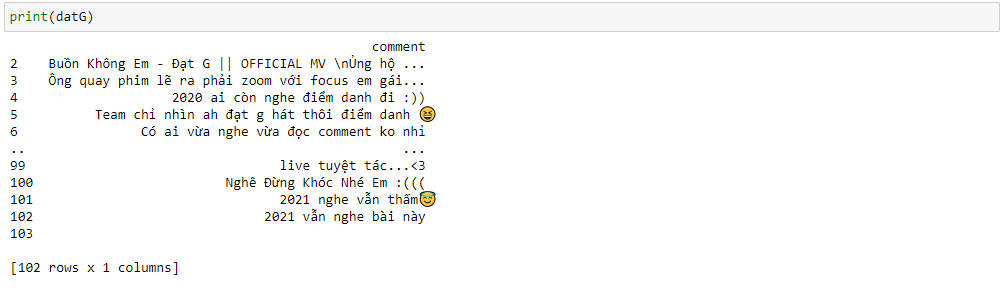
Tiếp đến ta sẽ làm mịn dữ liệu:

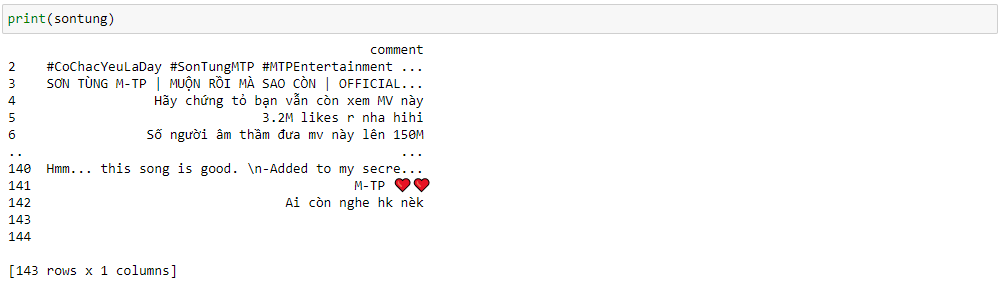


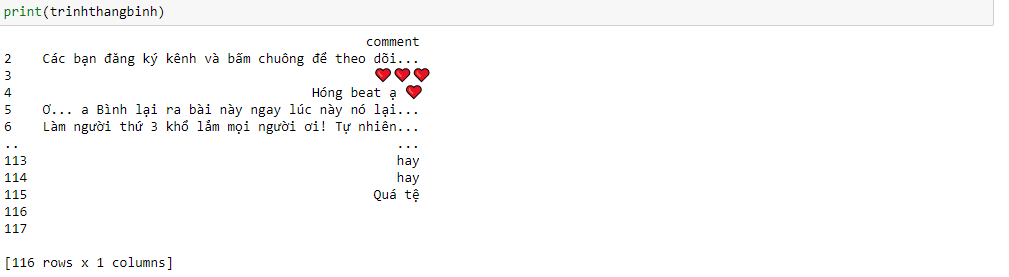
Sau đó sẽ gọi 3 tập dữ liệu làm mịn chúng



Chúng ta có thể nhìn thấy chúng:



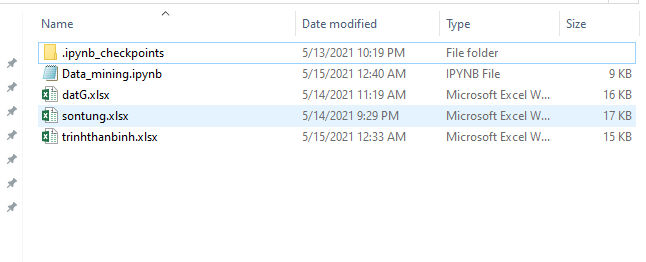




Tiếpđến chúng ta sẽ xuất chúng ra 3 file excel



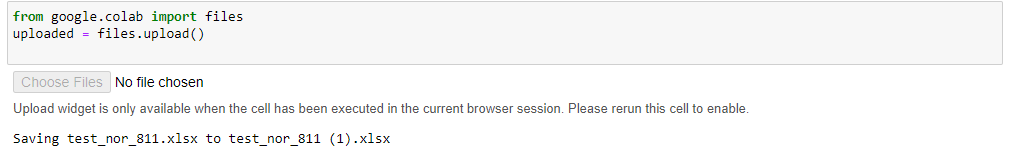
Khi đó ta sẽ có 3 file excel đã xuất ra tương ứng với từng tệp dữ liệu:

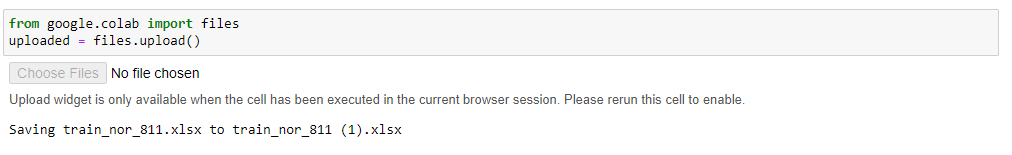


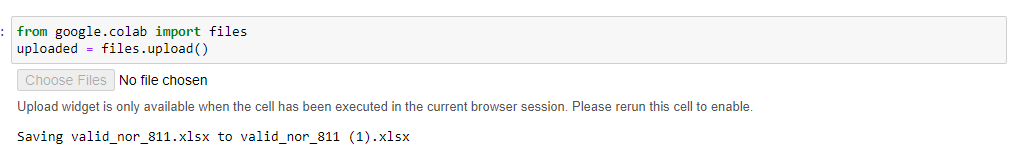
* 1. Task2:

Đầu tiên ta cần import các thư viện và sau đó upload file lên:

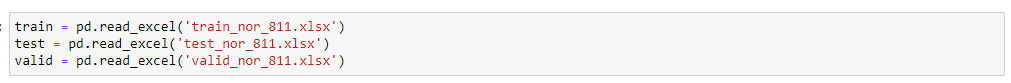




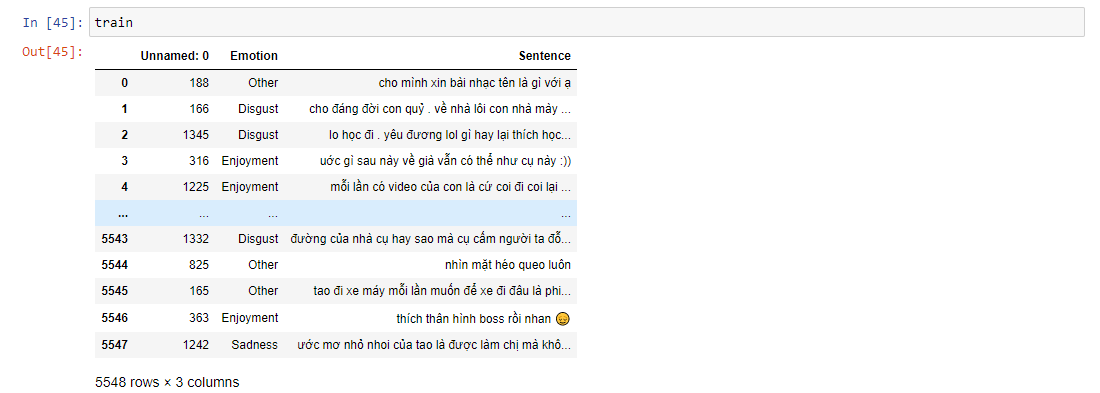


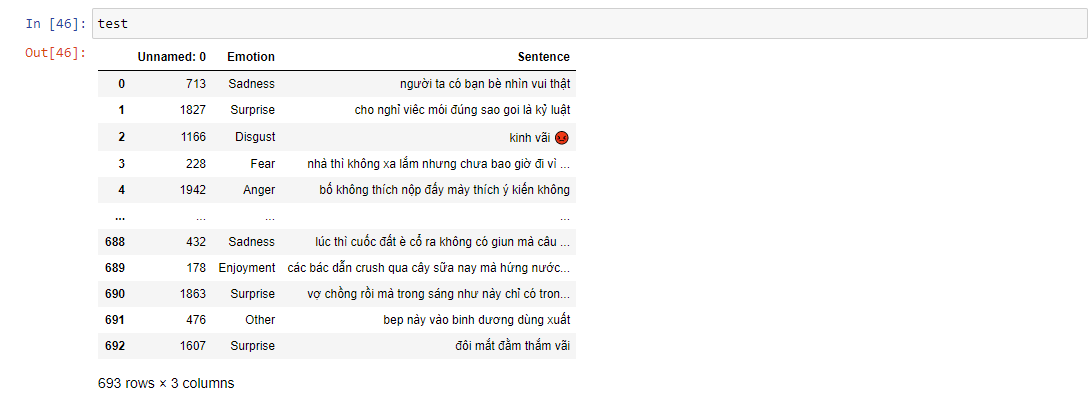


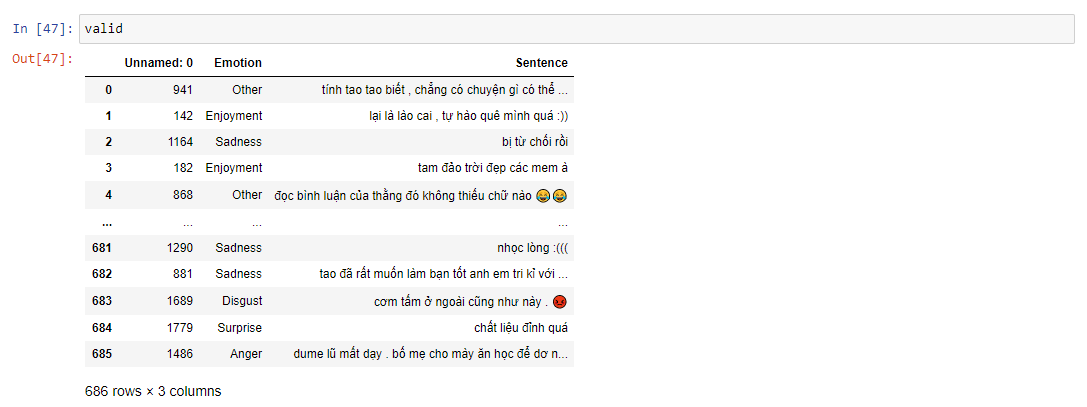
Ta sẽ đọc dữ liệu của 3 file train test valid:



Và ta sẽ print nó ra:







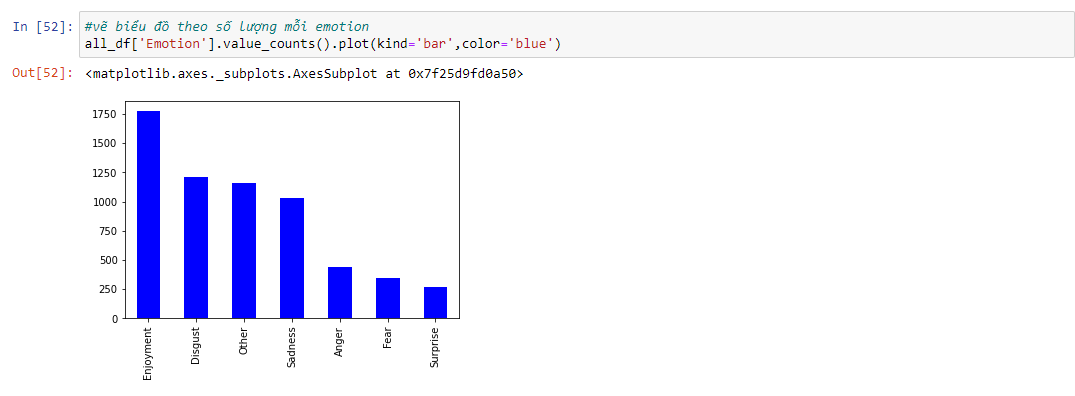
Tiếp theo ta sẽ gộp 2 file train và valid lại:



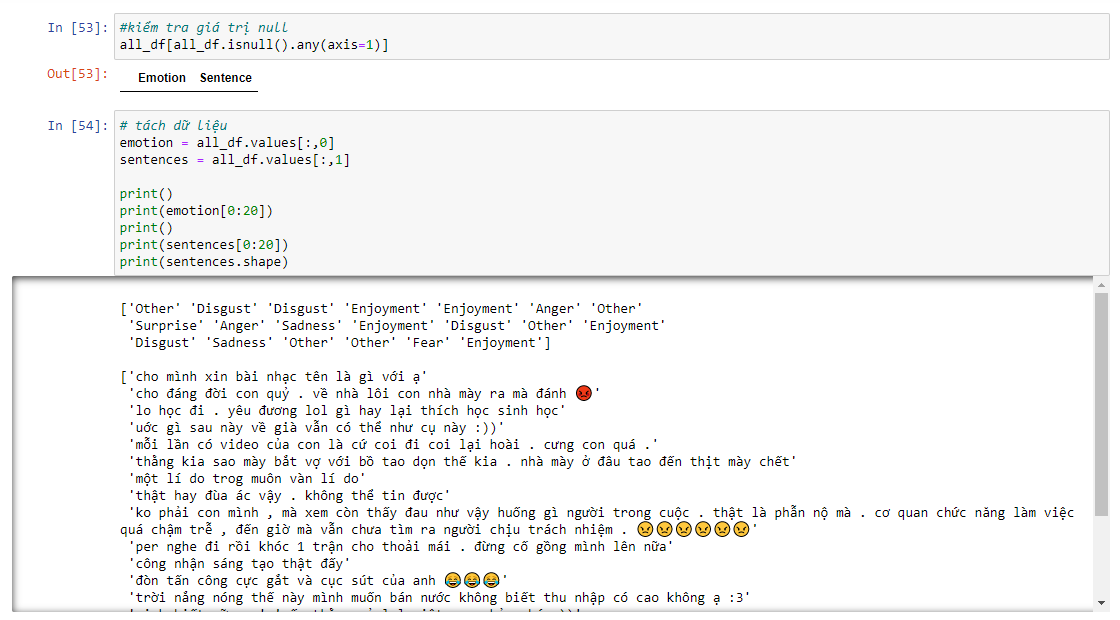
Sau đó ta sẽ bỏ các cột dư thừa và lấy lại 2 cột cuối đặt tên cho nó là emotion và sentence :



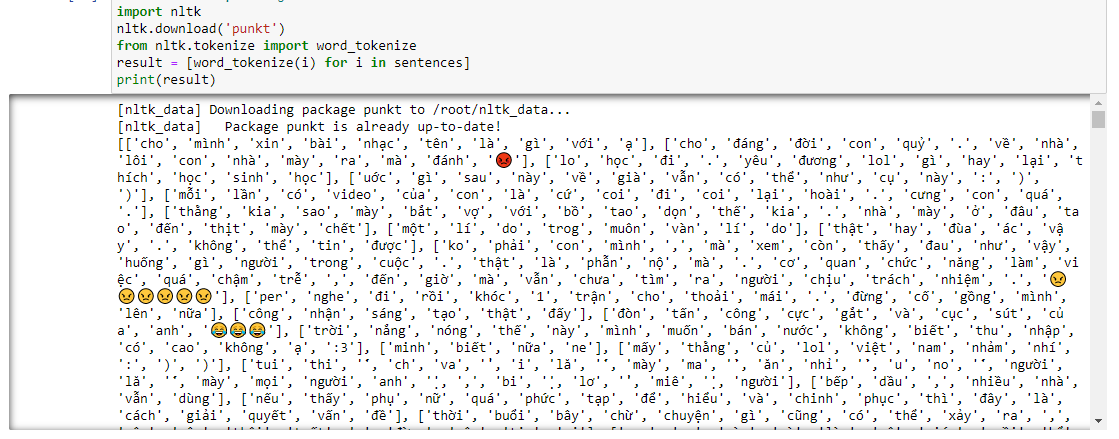
Ta sẽ vẽ biểu đồ theo số lượng mỗi emotion:



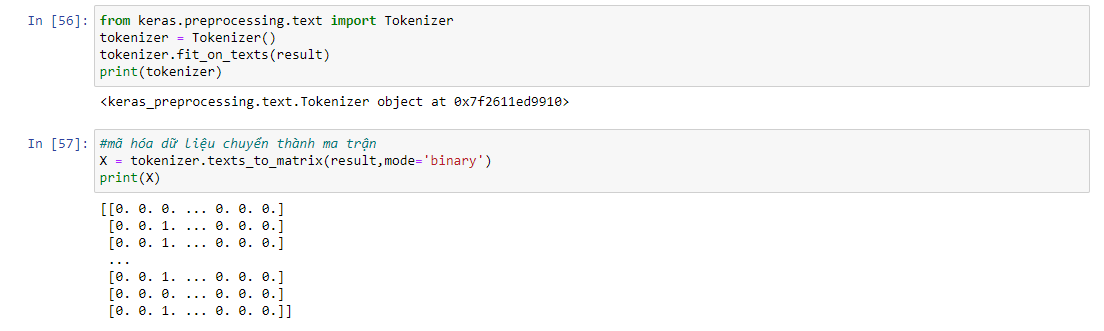
Tiếp theo ta sẽ kiểm tra giá trị null và tách dữ liệu của chúng ra:



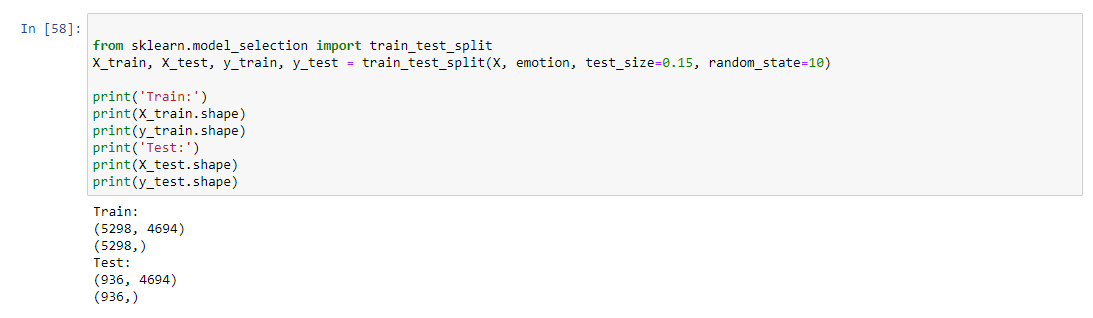
Sau đó ta sẽ trích xuất đặc trưng mỗi câu gồm nhiều từ:



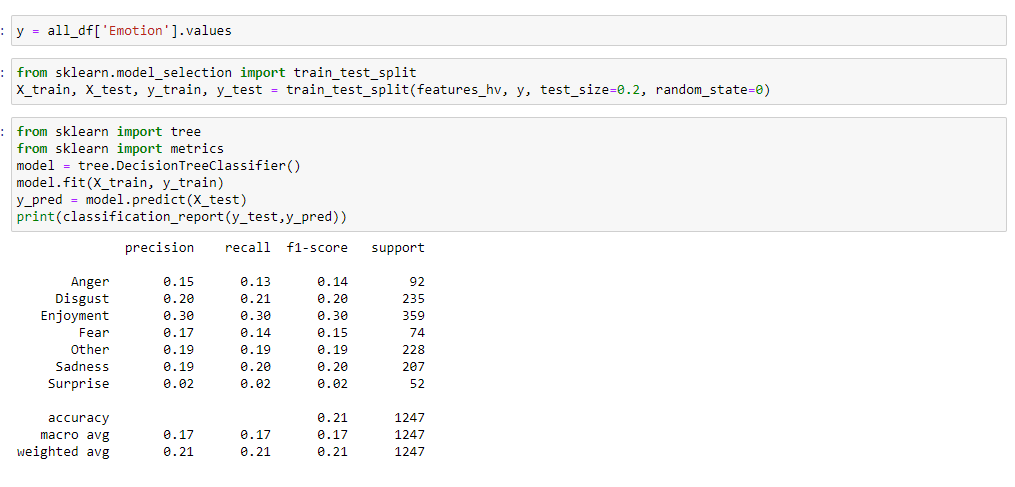
Ta sẽ token nó đi và chuyển thành ma trận:



Sau đó ta sẽ tách tập train và test:



Tiếp theo ta sử dụng Sklearn: ở đây test là 80% dữ liệu đầu train là 20% dữ liệu lại và chúng ta sẽ dự đoán cột Emotion

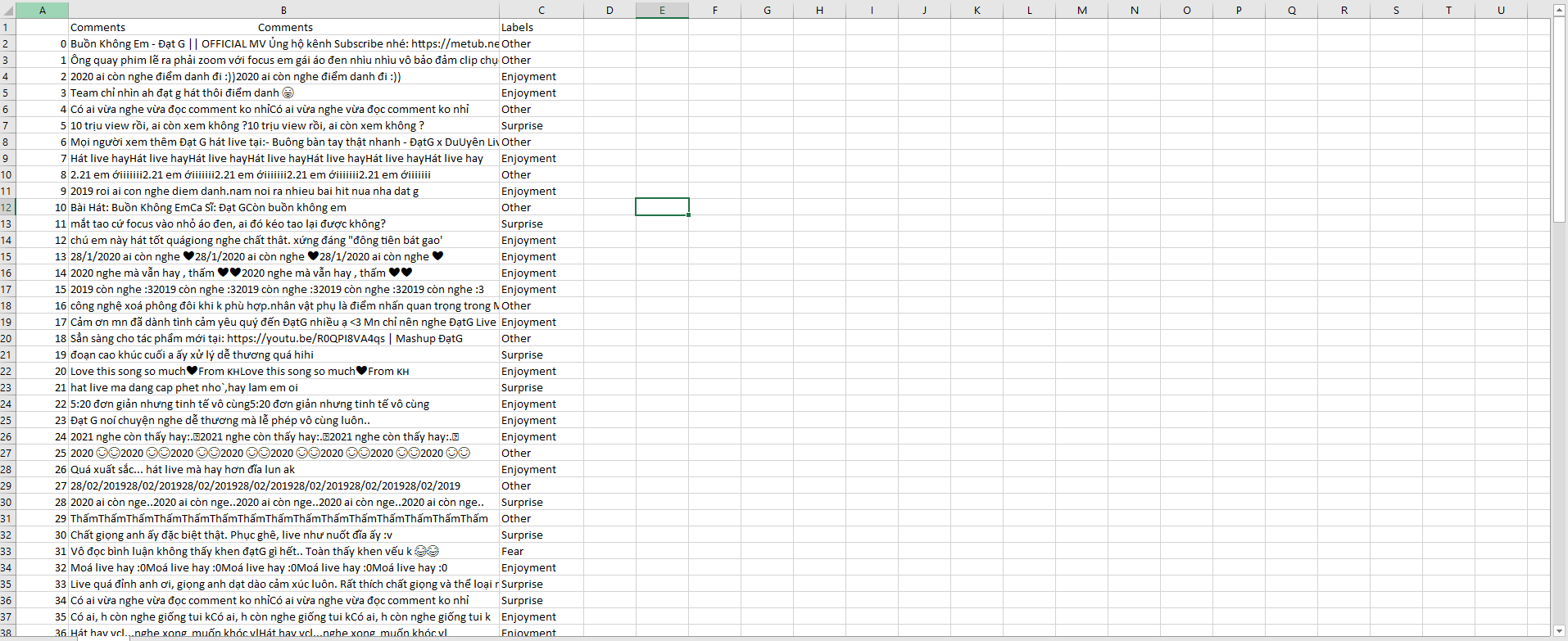


CHƯƠNG 3

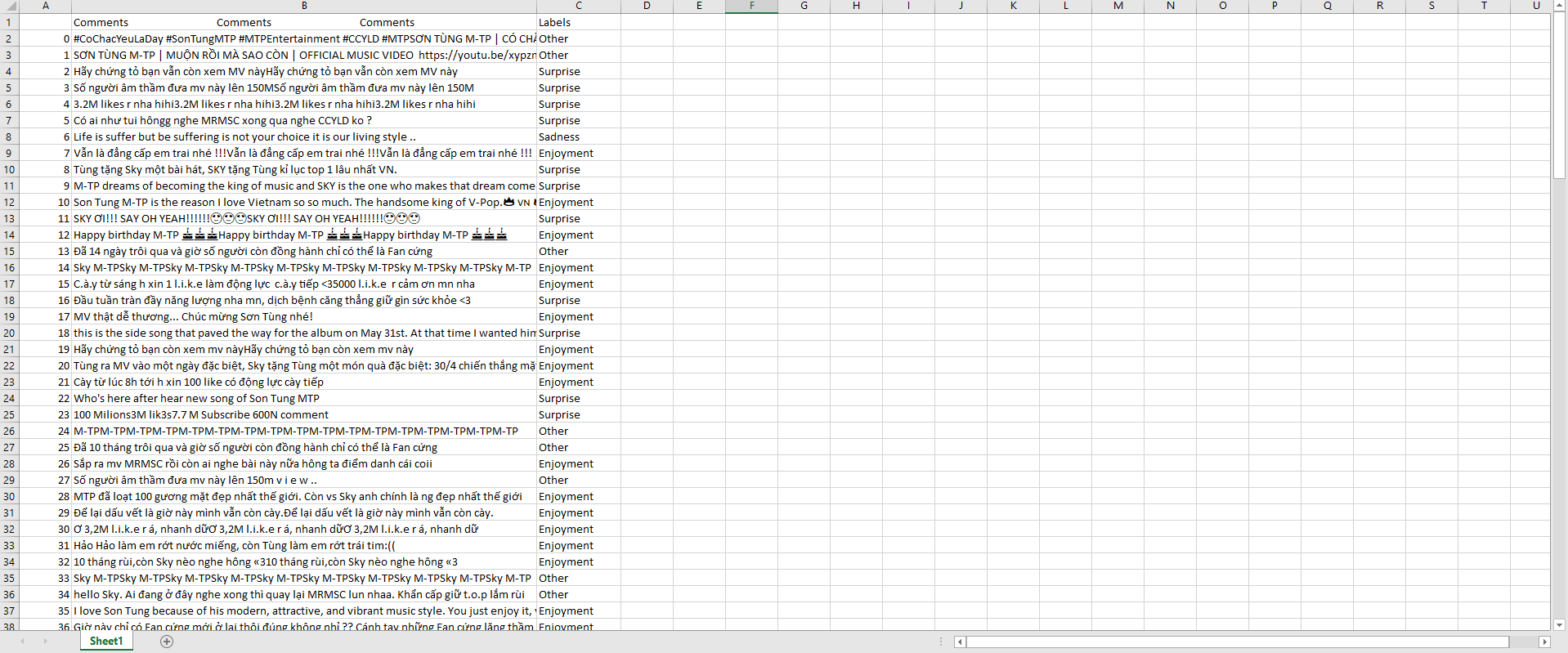
KẾT QUẢ

3.1 Task1:

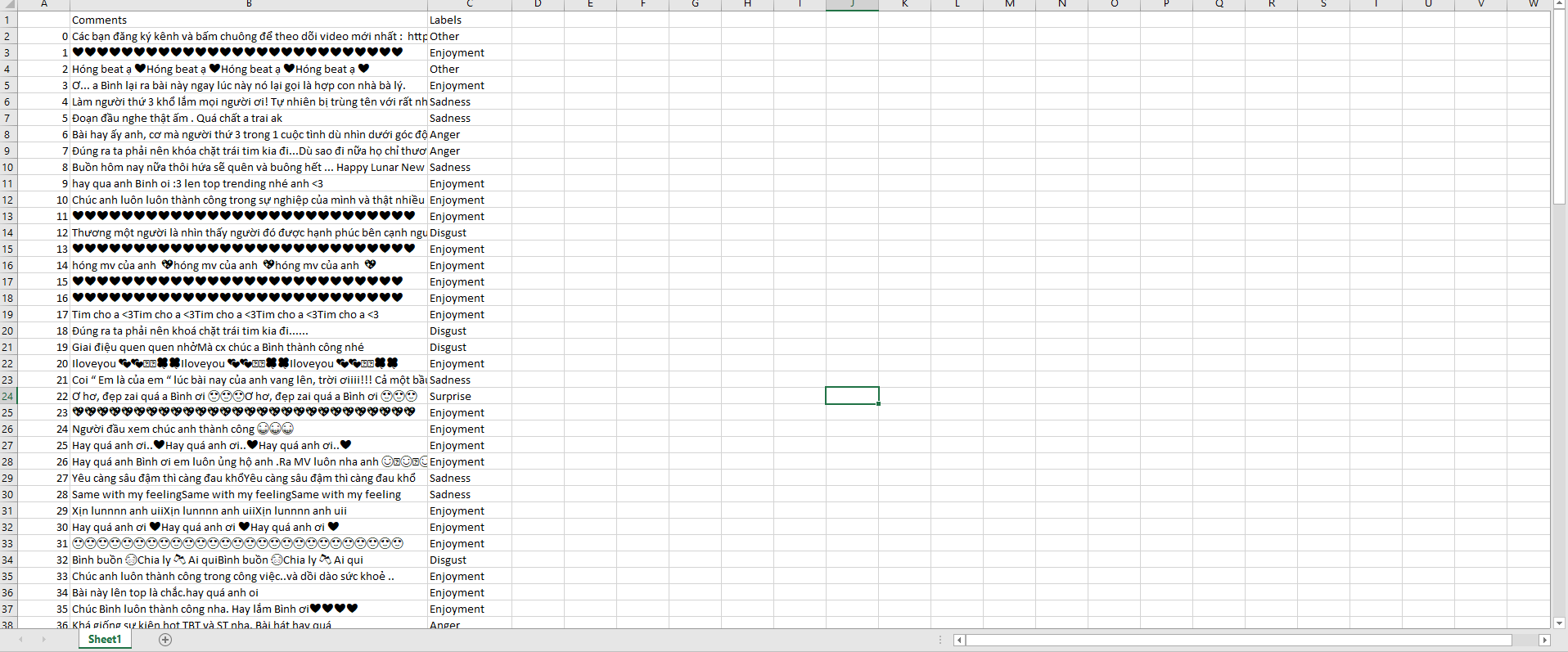
Đầu tiên sẽ là file datG.xlsx:



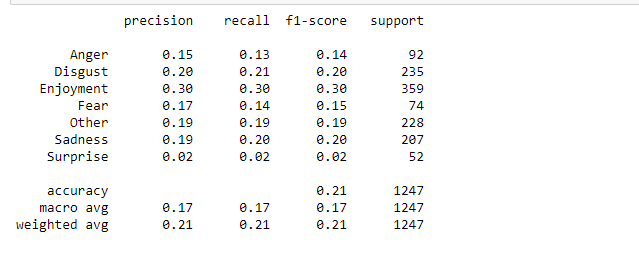
Và đây là file songtung.xslx:



Đây là file trinhthanhbinh.xlsx:



3.2 Task2:



**TÀI LIỆU THAM KHẢO**

1. Quách Ngọc Ân (1992), “Nhìn lại hai năm phát triển lúa lai”, *Di tuyền học ứng dụng*, 98(1), tr. 10-16.