TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DATA MINING AND KNOWLEDGE DISCOVERY**

**FINAL PROJECT**

*Người hướng dẫn:* **LÊ CUNG TƯỞNG**

*Người thực hiện*: **NGUYỄN TRẦN NHẬT ĐỨC – 518H0609**

**NGUYỄN KHẮC MINH LUÂN – 518H0398**

Lớp **: 18H50302**

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**



**DATA MINING AND KNOWLEDGE DISCOVERY**

**FINAL PROJECT**

Người hướng dẫn: **LÊ CUNG TƯỞNG**

Người thực hiện: **NGUYỄN TRẦN NHẬT ĐỨC**

**NGUYỄN KHẮC MINH LUÂN**

Lớp **: 18H50302**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Chúng em xin bày tỏ lòng biết ơn sâu sắc tới thầy Lê Cung Tưởng vì những ý kiến ​​đóng góp có giá trị và mang tính xây dựng trong quá trình lập kế hoạch và phát triển dự án thuật toán thiết kế và phân tích này. Những lời khuyên của thầy Lê Cung Tưởng đã giúp ích rất nhiều cho việc làm bài tiểu luận cuối kỳ này.

Chúng em cũng muốn gửi lời cảm ơn đến ban Giám hiệu nhà trường, tất cả những người đã cung cấp cho chúng em các cơ sở vật chất cần thiết và các điều kiện dẫn điện cho dự án.

Xin chân thành cảm ơn với tất cả sự kính trọng của chúng em.

**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 Thầy Lê Cung Tưởng. 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)*

*Đã ký*

*Nguyễn Trần Nhật Đức*

*Đã ký*

*Nguyễn Khắc Minh Luân*

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)

Table of Contents

[TASK 1. Collecting data 6](#_Toc72067654)

[1.1. Việc cần làm: 6](#_Toc72067655)

[1.2. Code: 6](#_Toc72067656)

[TASK 2. Emotion Recognition for Vietnamese Social Media Text 9](#_Toc72067657)

[1.1. Việc cần làm: 9](#_Toc72067658)

[1.2. Giải thích Code và Output: 10](#_Toc72067659)

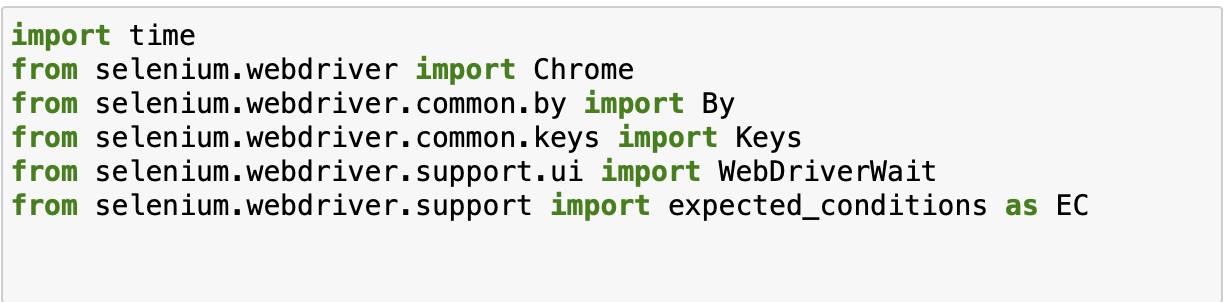
[TASK 3. Applying the trained model in Task 2 to three datasets in Task 1 24](#_Toc72067660)

[1.1. Việc cần làm: 24](#_Toc72067661)

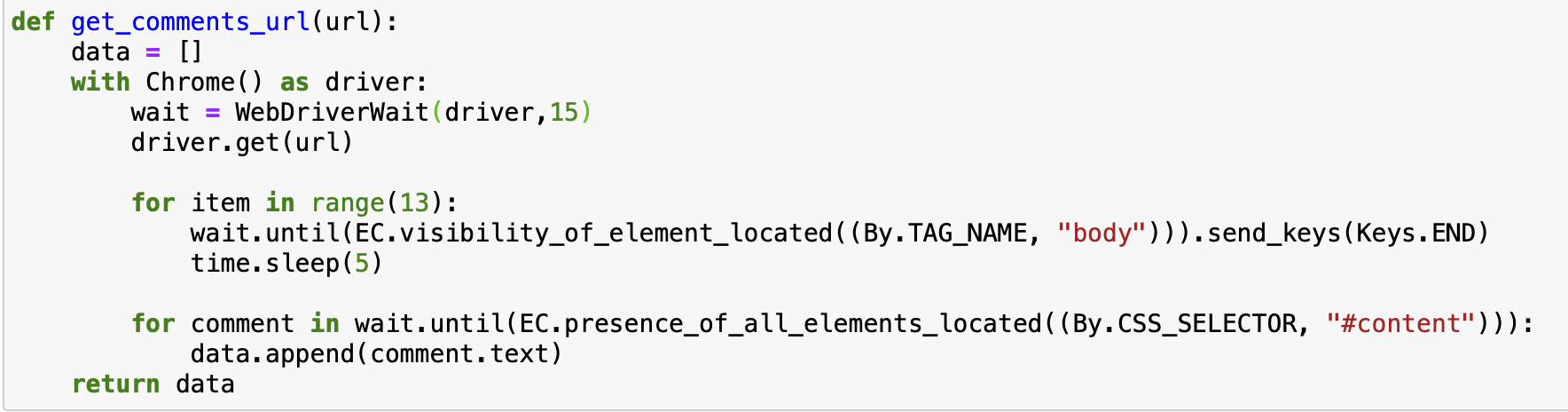
[1.2. Giải thích code và Output: 24](#_Toc72067662)

**TASK 1. Collecting data**

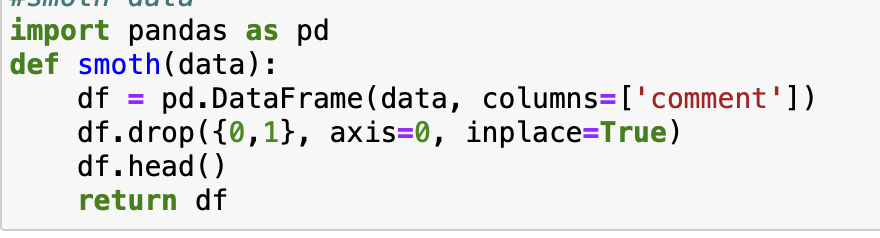
* 1. **Việc cần làm:**
* Ở Task 1 thực hiện lấy các Comments trên Youtube để làm dữ liệu. Lấy 100 Comments mỗi Video trên Youtube, tổng cộng là 300 Comments. Sau đó gắn Tag(Emotion) cho mỗi câu. Có 7 loại cảm xúc : Disgust (chán ghét), Enjoyment (thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi), Other (khác).
  1. **Code:**
* Bước đầu là import và sử dụng thư viện Selenium có kết hợp API để lấy Comments từ Youtube



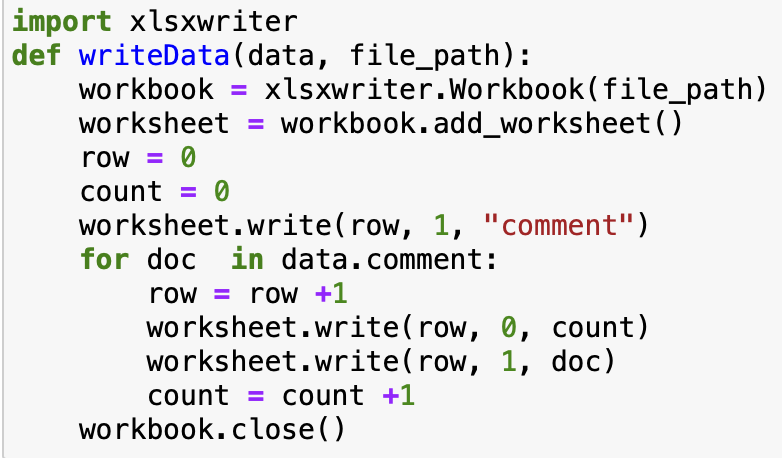
* Tạo hàm Def để có thể dùng Crawls dữ liệu nhiều lần



* Tạo DataFrame chứa Comments Crawls được từ Youtube :

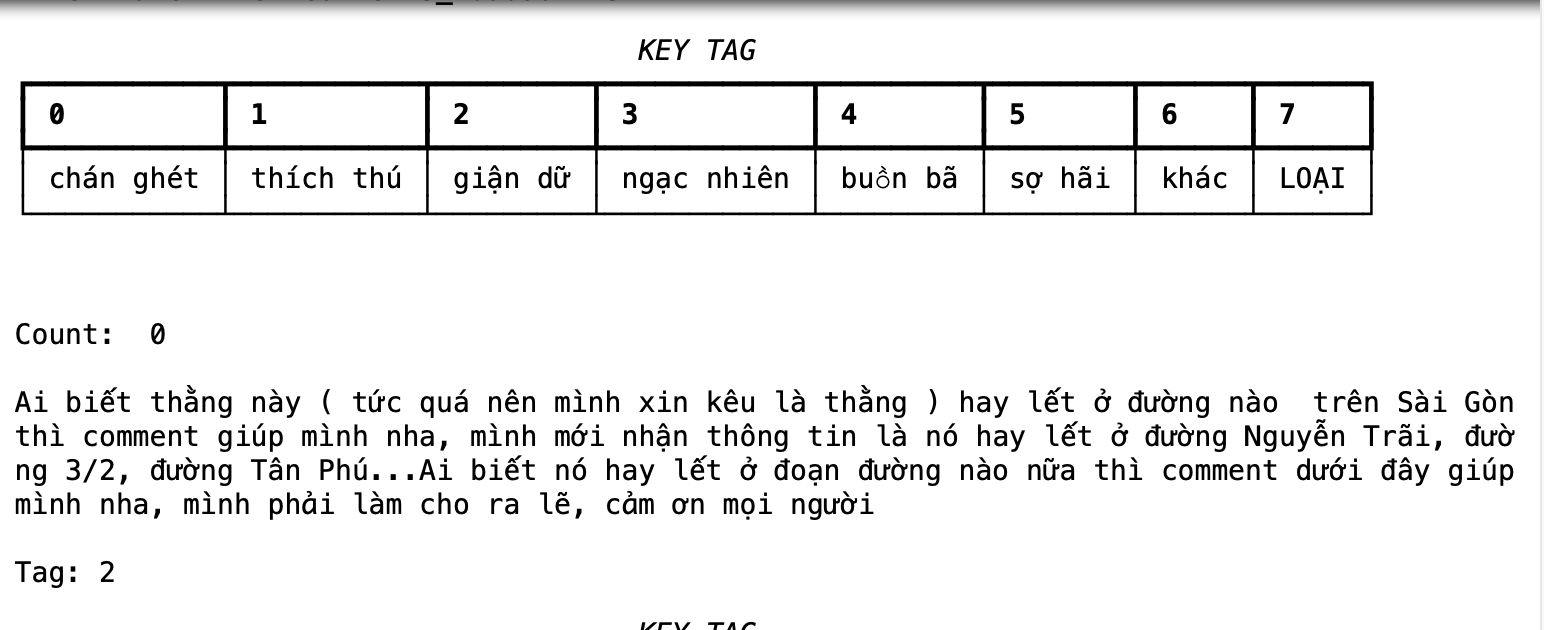


* Tạo hàm Def dùng để đưa Comments vào file excel:



* Sử dụng 1 code khác để gắn Tag(Emotion) cho các file dữ liệu: gắn tag theo số từ 0 đến 6 tương ứng với các Cảm xúc để tiết kiệm thời gian, sau đó chuyển từ số sang Cảm xúc và ghi vào file excel

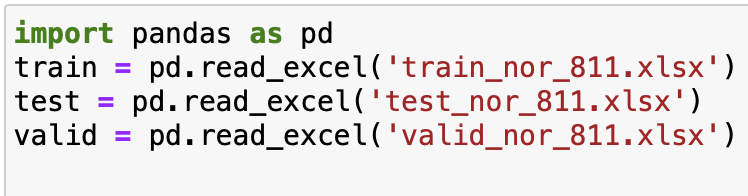




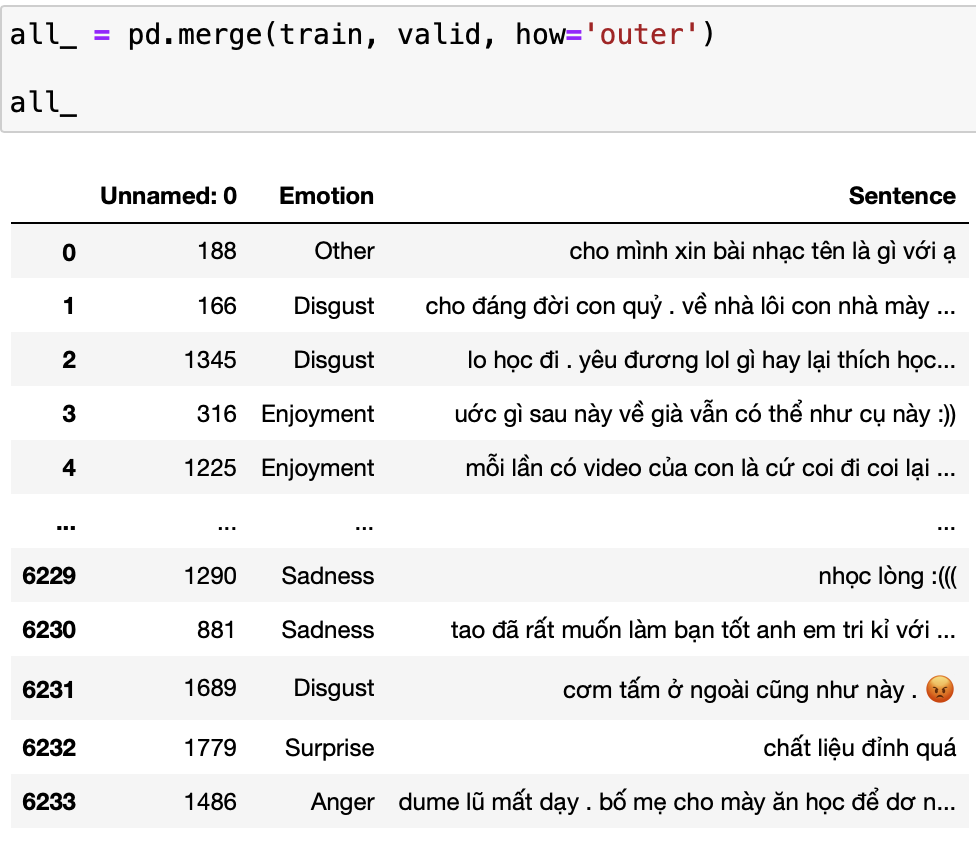


**TASK 2. Emotion Recognition for Vietnamese Social Media Text**

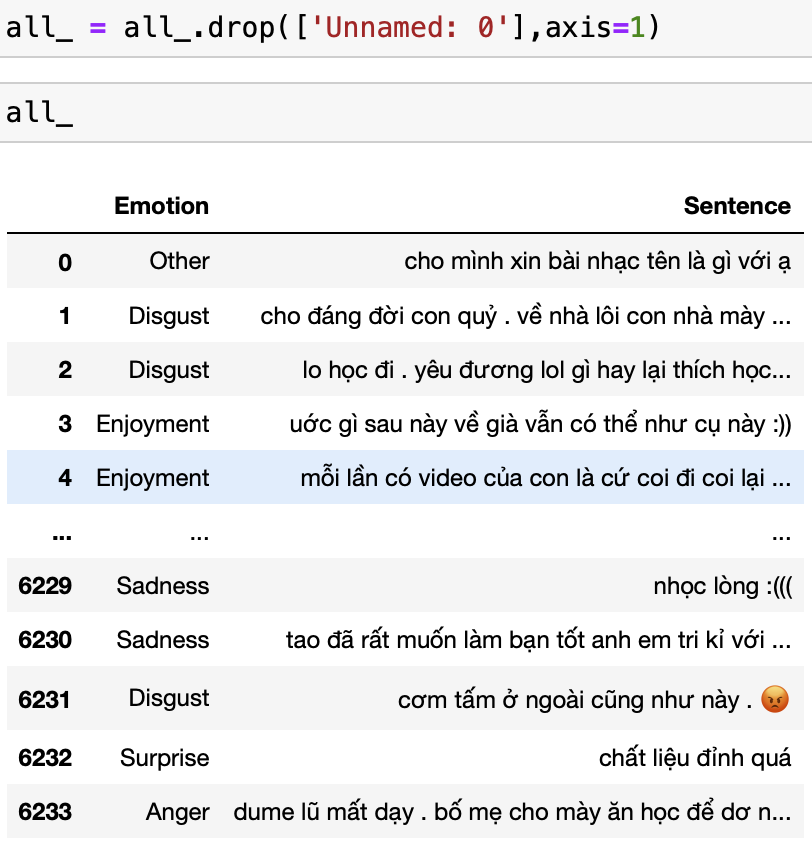
* 1. **Việc cần làm:**
* Sử dụng tệp train và valid để tạo tập dữ liệu huấn luyện nhận biết cảm xúc của một câu, thuộc loại classification. Có thể sử dụng bất kỳ model classifer nào mong muốn
* Sau khi kết thúc huấn luyện model thì sẽ áp dụng lên tập Test và đưa ra độ chính xác so với thực tế.
  1. **Giải thích Code và Output:**
* Ở phần code, chúng em làm dữ liệu huấn luyện dựa trên 2 thư viện là Keras và Sklearn kết hợp với thư viện tách từ Tiếng Việt là Underthesea và Pyvi để so sánh cái nào sẽ cho Accuracy tốt hơn.
* Đọc dữ liệu các file với pandas:



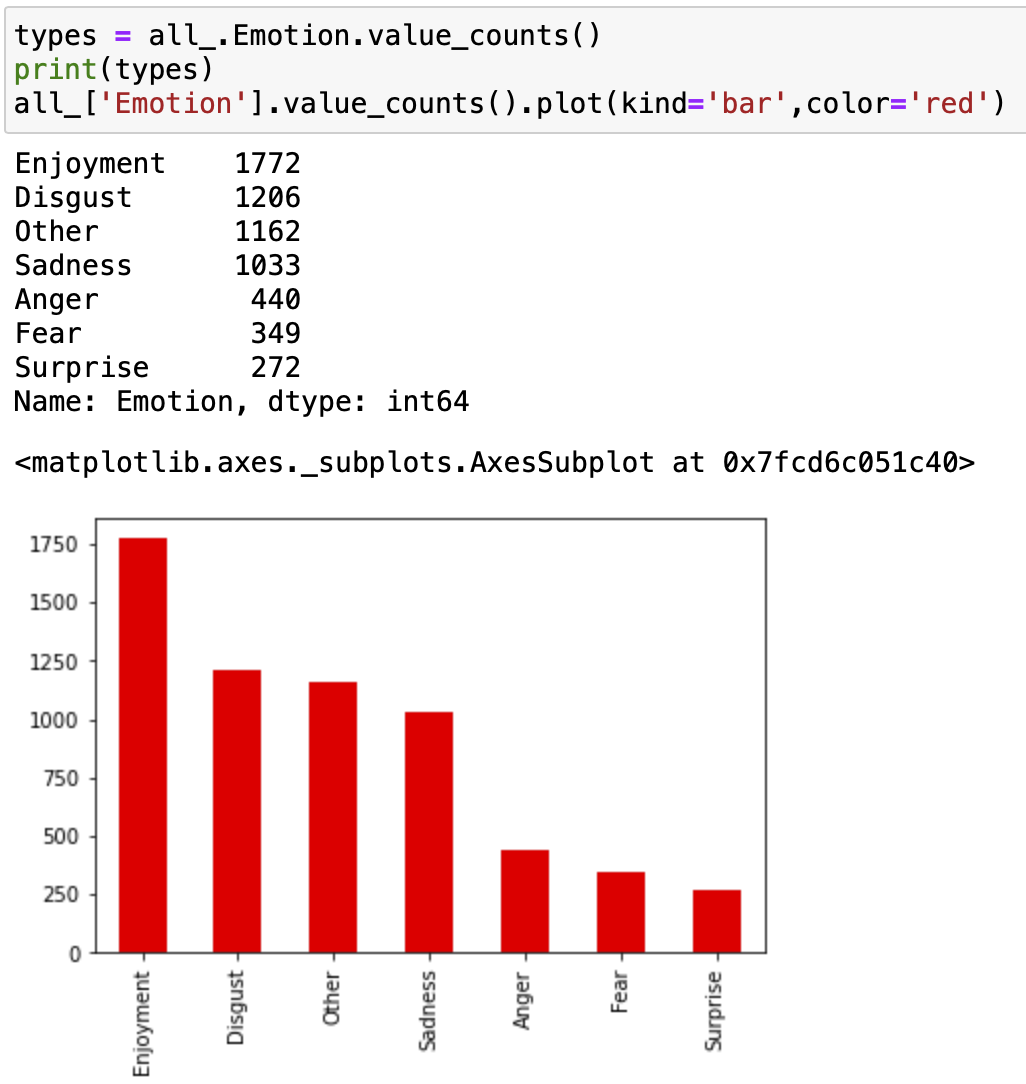
* Sử dụng lệnh **Merge** để nối 2 file excel train và valid với nhau như một file Train lớn để làm dữ liệu huấn luyện:



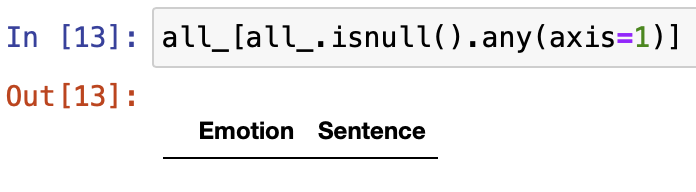
* Sử dụng lệnh **drop** để loại bỏ cột dư thừa:



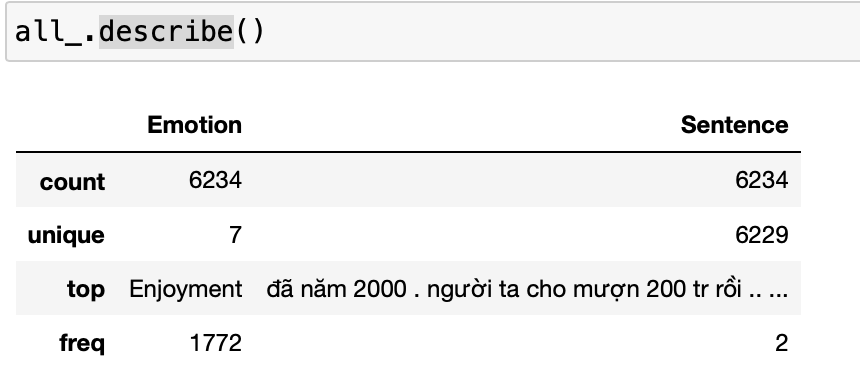
* Vẽ và in ra số lượng Cảm xúc trong bộ dữ liệu:



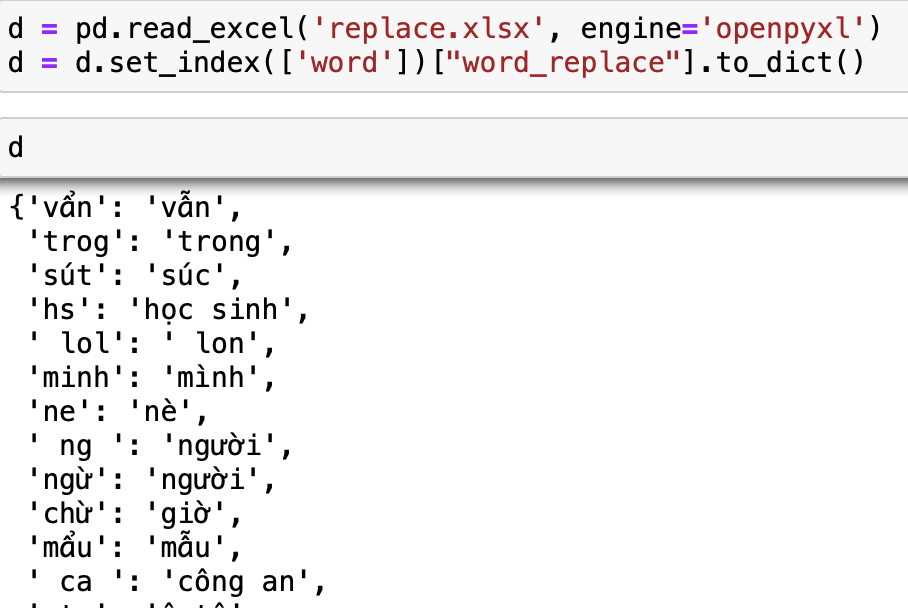
* Kiểm tra dữ liệu Null để tránh lỗi dữ liệu dùng lệnh **isnull()**



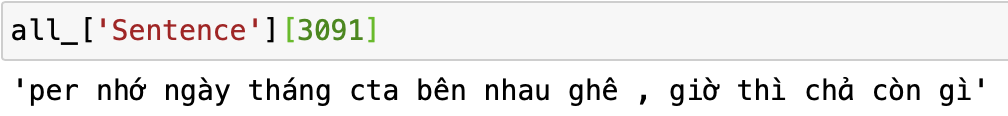
* Miêu tả dữ liệu dùng **Descibe:**



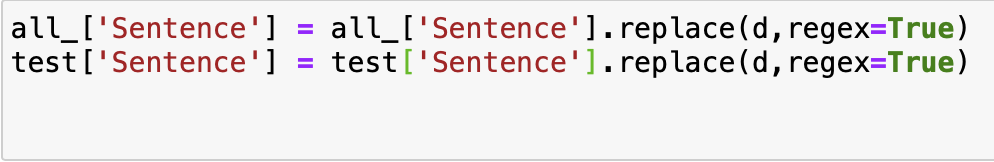
* Tạo 1 file excel dùng để sửa lỗi chính tả trong tập dữ liệu để tăng độ chính xác khi huấn luyện dữ liệu. Đọc file và chuyển dữ liệu về dạng dict và áp dụng lệnh **replace** sửa lỗi chính tả:



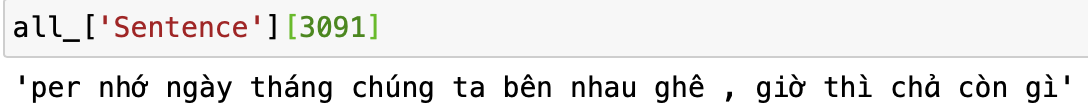
* Dữ liệu trước khi áp dụng sửa chính tả:



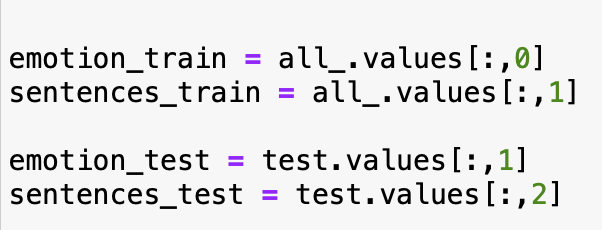
* Áp dụng sửa lỗi chính tả vào dữ liệu:



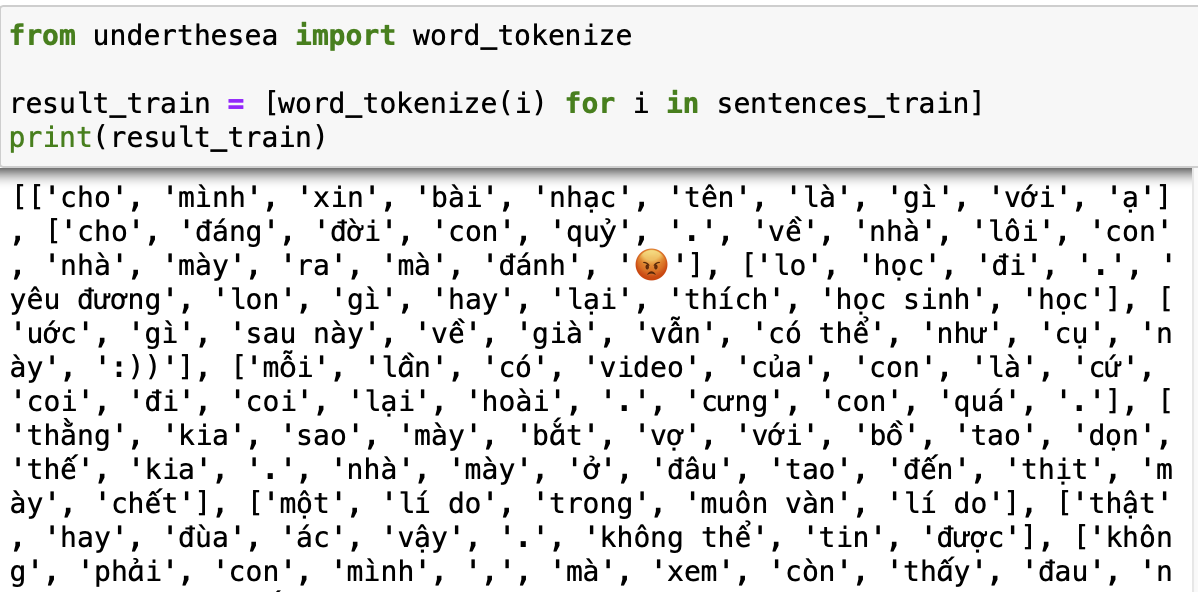
* Sau khi áp dụng:

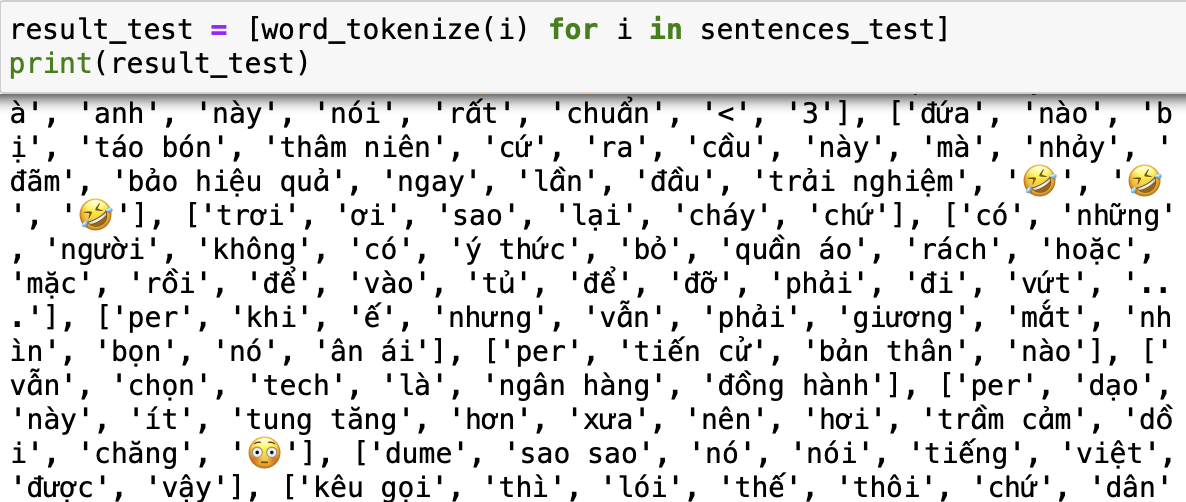


* Lấy dữ liệu từ tập Train và tập Test để dùng cho huấn luyện:

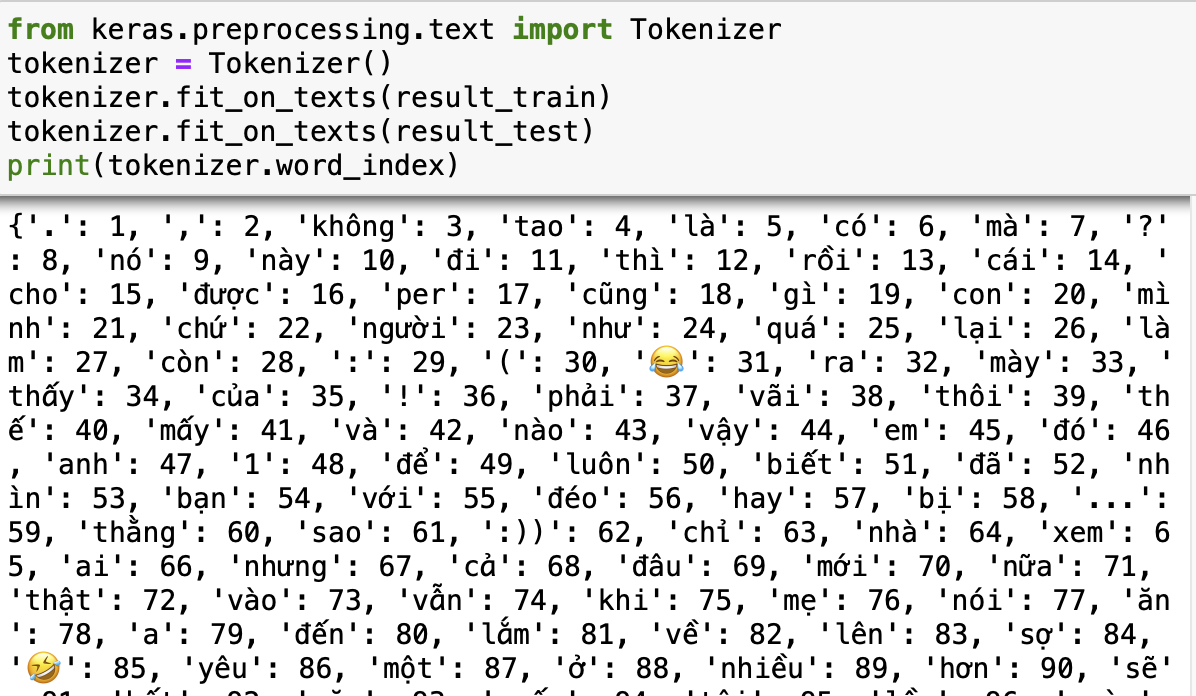


* Dùng thư viện **Underthesea** để tách từ Tiếng Việt cho các câu trong tập dữ liệu Train, Test:

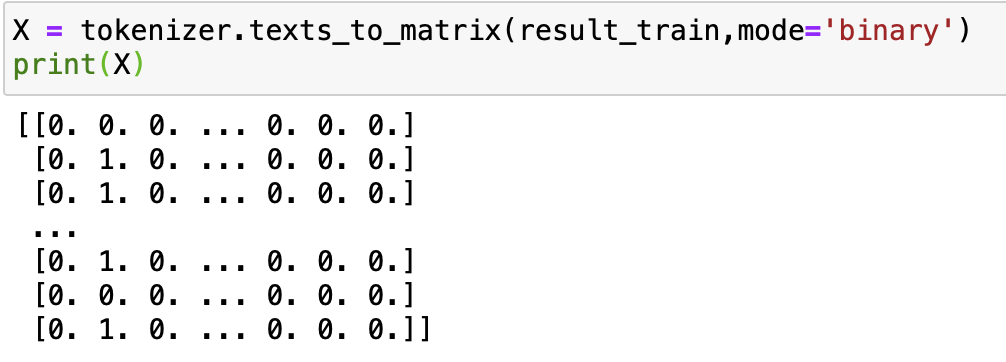


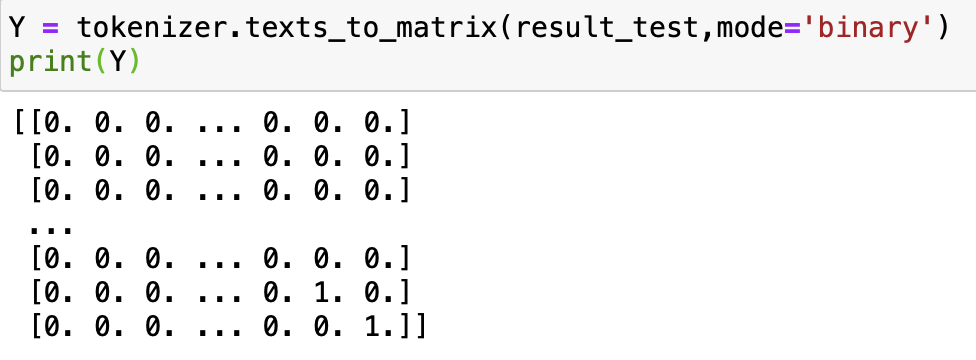


* Dùng bộ phân tích từ vựng **Tokenizer** của Keras cho dữ liệu Train và Test:

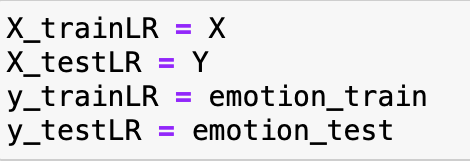


* Chuyển đổi dữ liệu về dạng ma trận sử dụng **mode Binary**

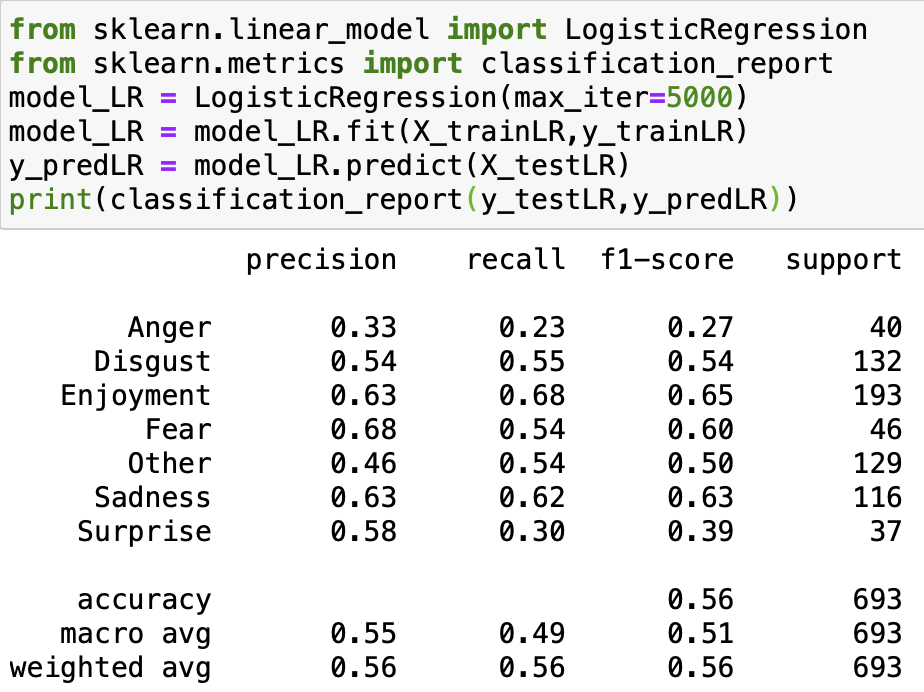




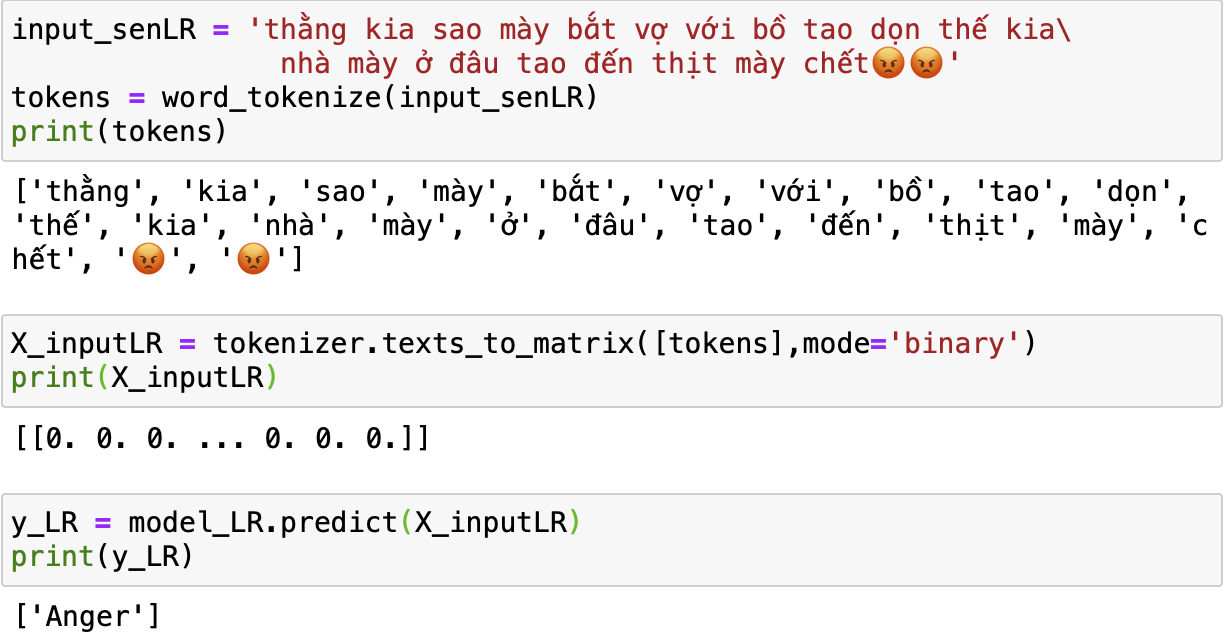
* Chia train test dữ liệu:



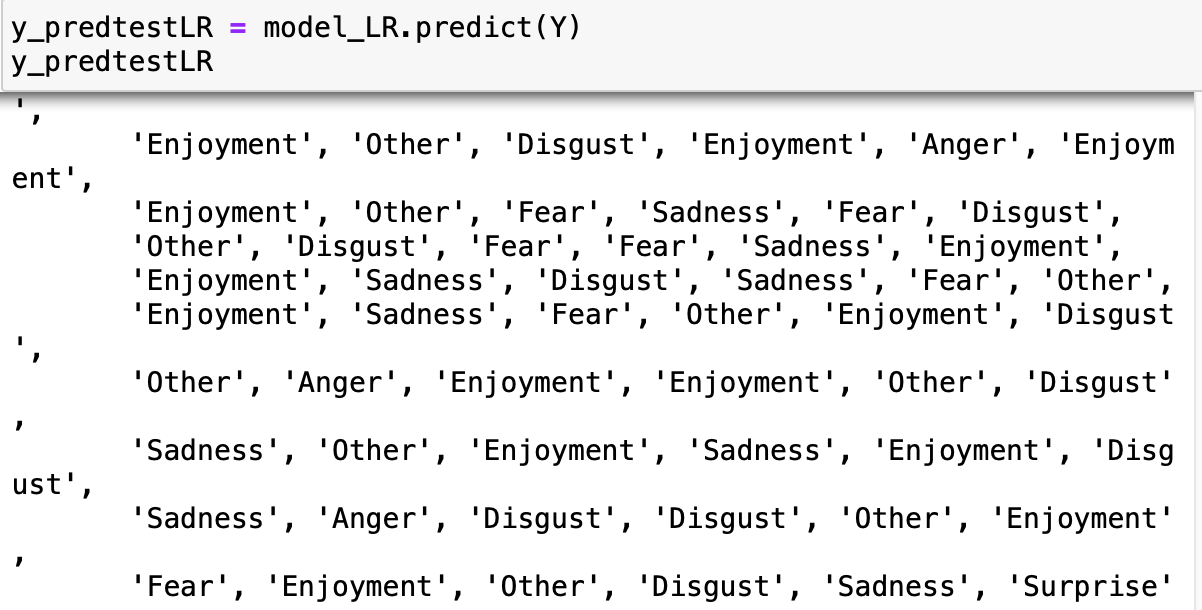
* Sử dụng mô hình **LogisticRegression** để huấn luyện dữ liệu:



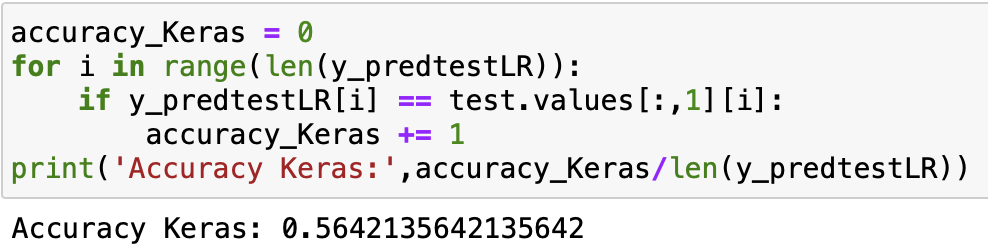
* Lí do Accuracy thấp vì nhiều lí do như: từ viết tắt, các emotions, nhiều từ sai chính tả chưa sửa hết
* Nhập 1 câu và dự đoán cảm xúc:



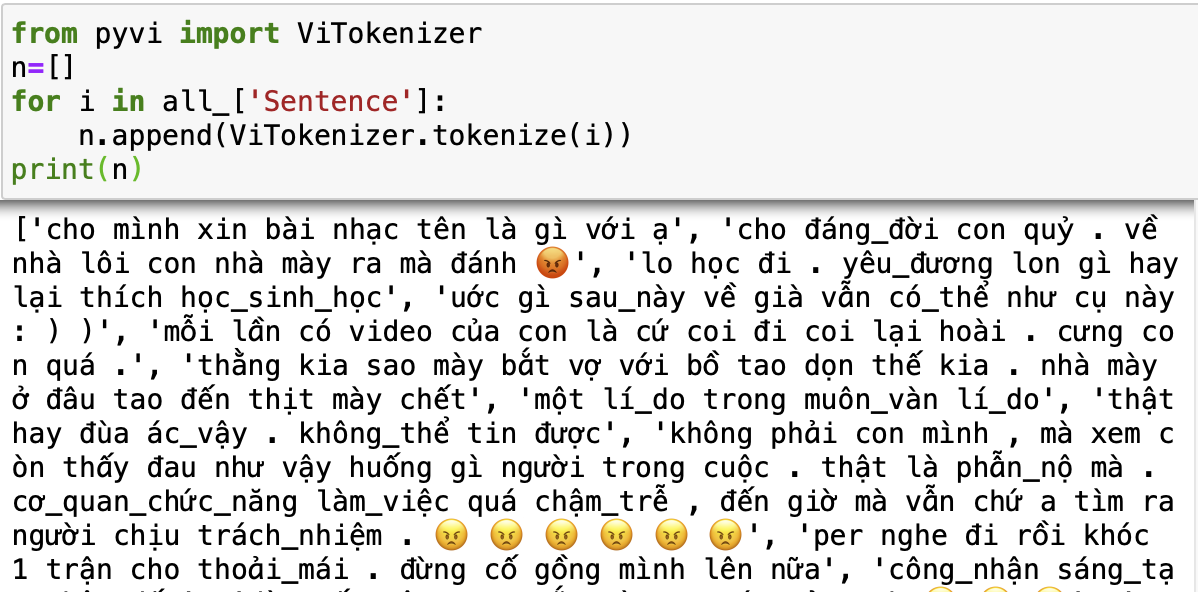
* Dùng model dự đoán cảm xúc cho các câu trong file Test:

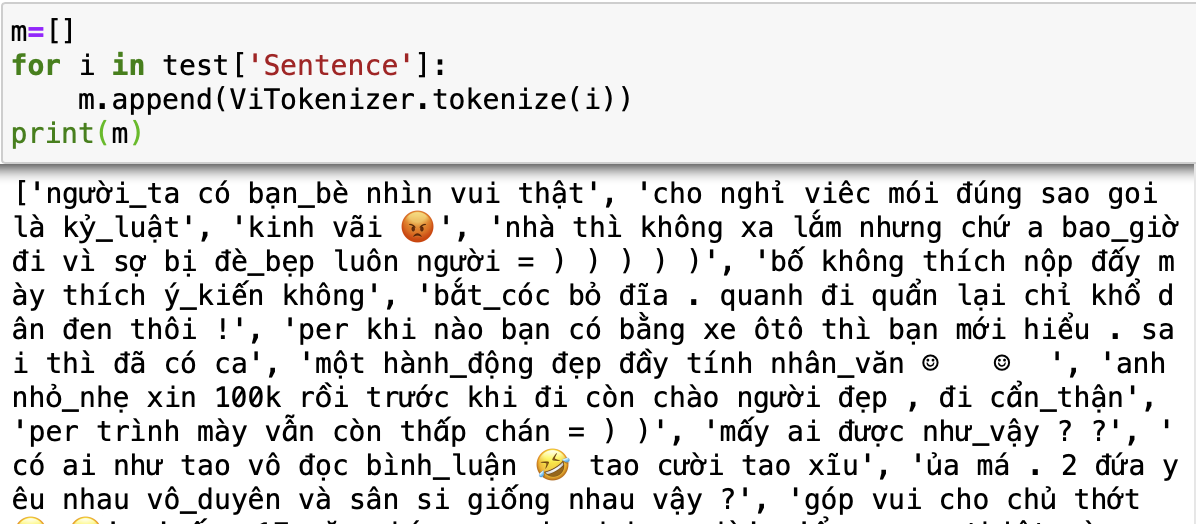


* Tính độ chính xác của mô hình **LogisticRegression** khi dự đoán cảm xúc so với dữ liệu file Test:

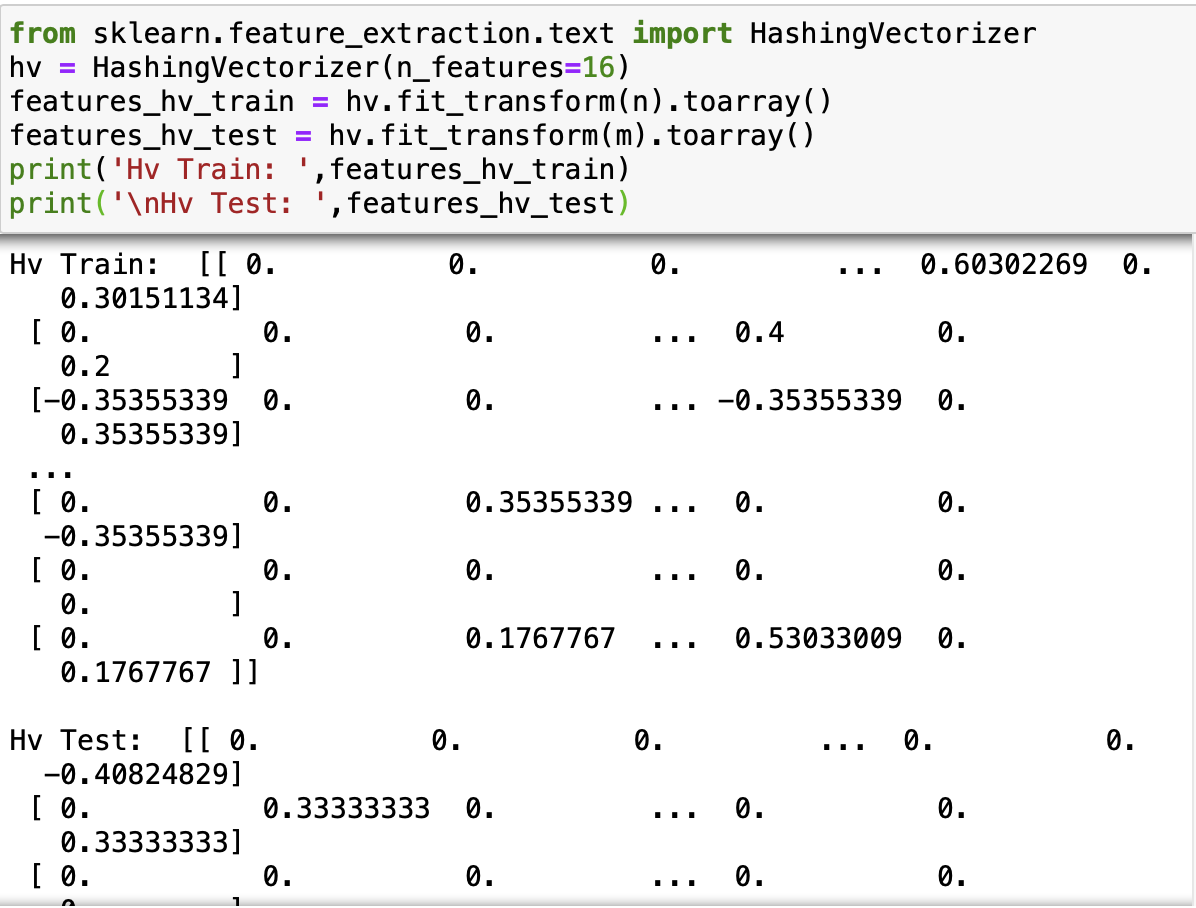


* **SỬ DỤNG SKLEARN**
* Áp dụng thư viện tách từ Tiếng Việt vào các câu trong tập dữ liệu Test và Train:

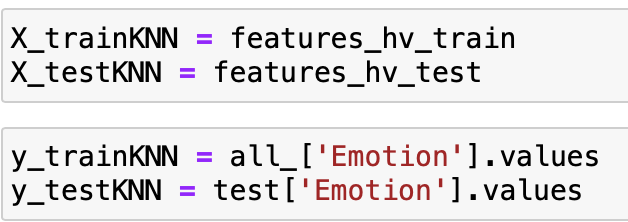




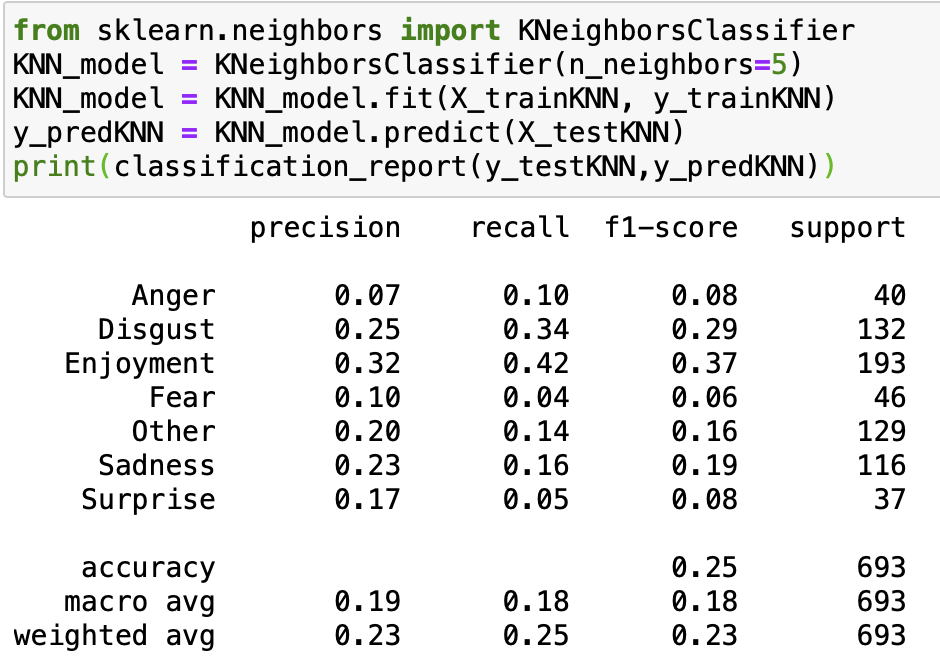
* Sử dụng HashingVectorizer cho cả bộ dữ liệu Train và Test vừa tách từ Tiếng Việt và chuyển chúng về dạng ma trận để đưa vào mô hình dự đoán:



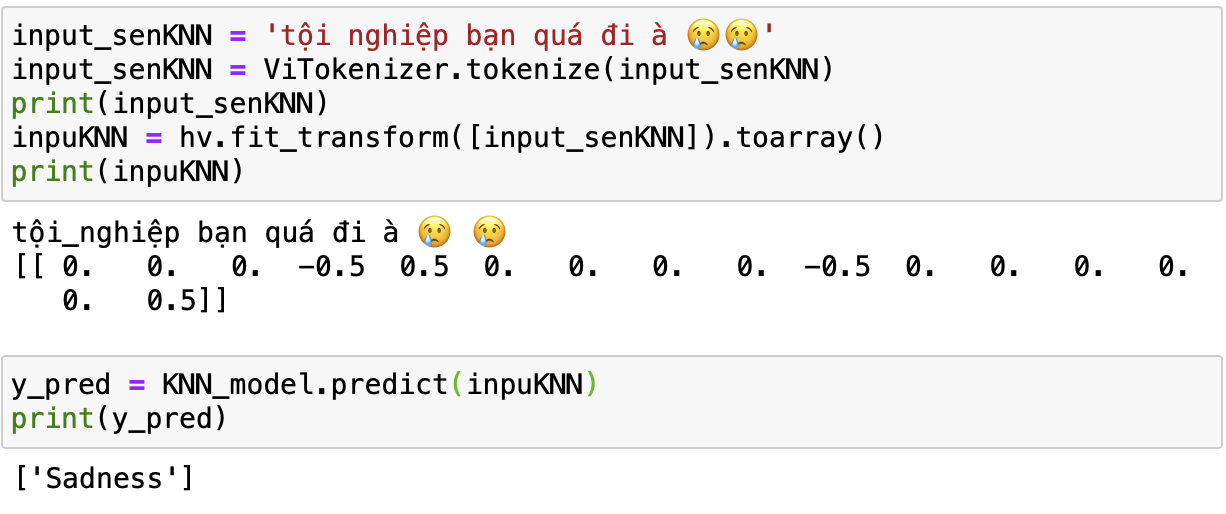
* Làm bộ dữ liệu Train, Test để chuẩn bị huấn luyện model:



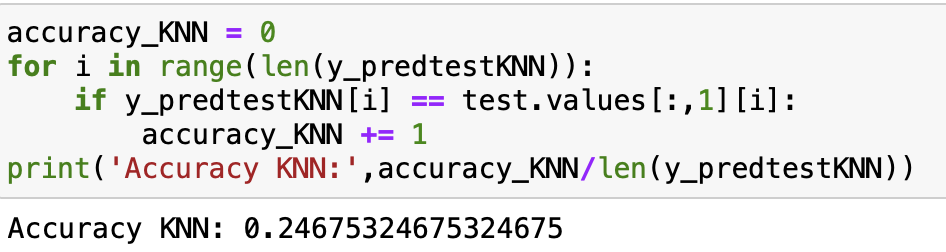
* Sử dụng Model KNN để làm dự đoán:



* Thử nhập câu và dự đoán cảm xúc của câu đó:

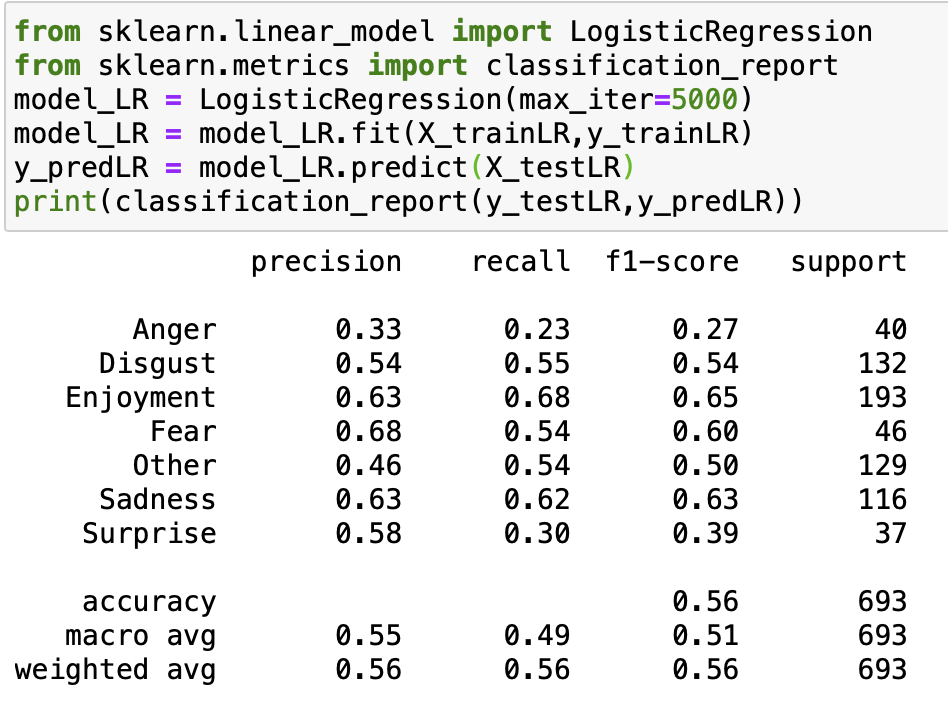


* Tính độ chính xác của Model KNN sử dụng thư viện Keras so với tập dữ liệu Test:



TASK 3. Applying the trained model in Task 2 to three datasets in Task 1

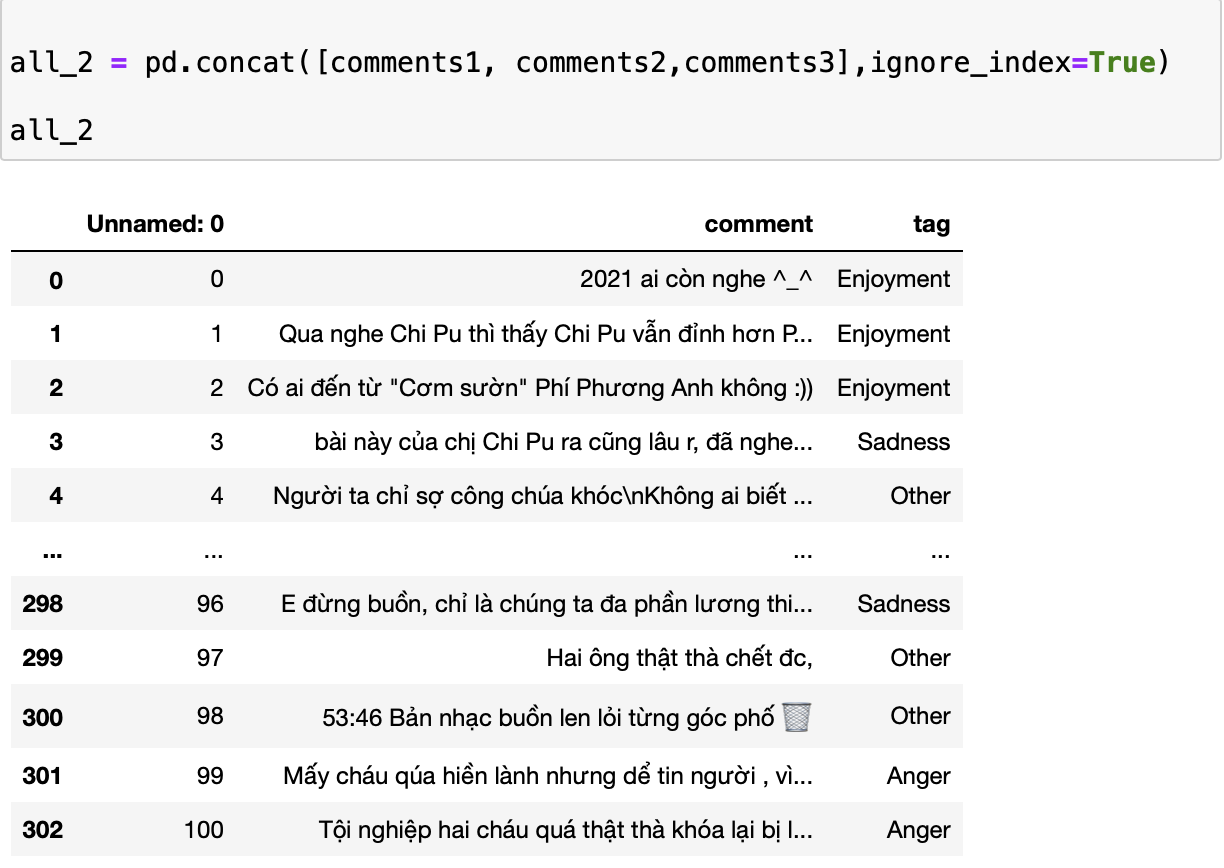
* 1. Việc cần làm:
* Sử dụng lại mô hình dự đoán của Task 2 để dự đoán cảm xúc các câu Comments Youtube lấy được ở Task 1. Sau đó tính toán độ chính xác dựa trên dữ liệu tự gắn
  1. Giải thích code và Output:
* Sử dụng lại mô hình dự đoán Cảm xúc LogisticRegression của Task 2 với độ chính xác của mô hình là 0.56 để dự đoán cho Task 3



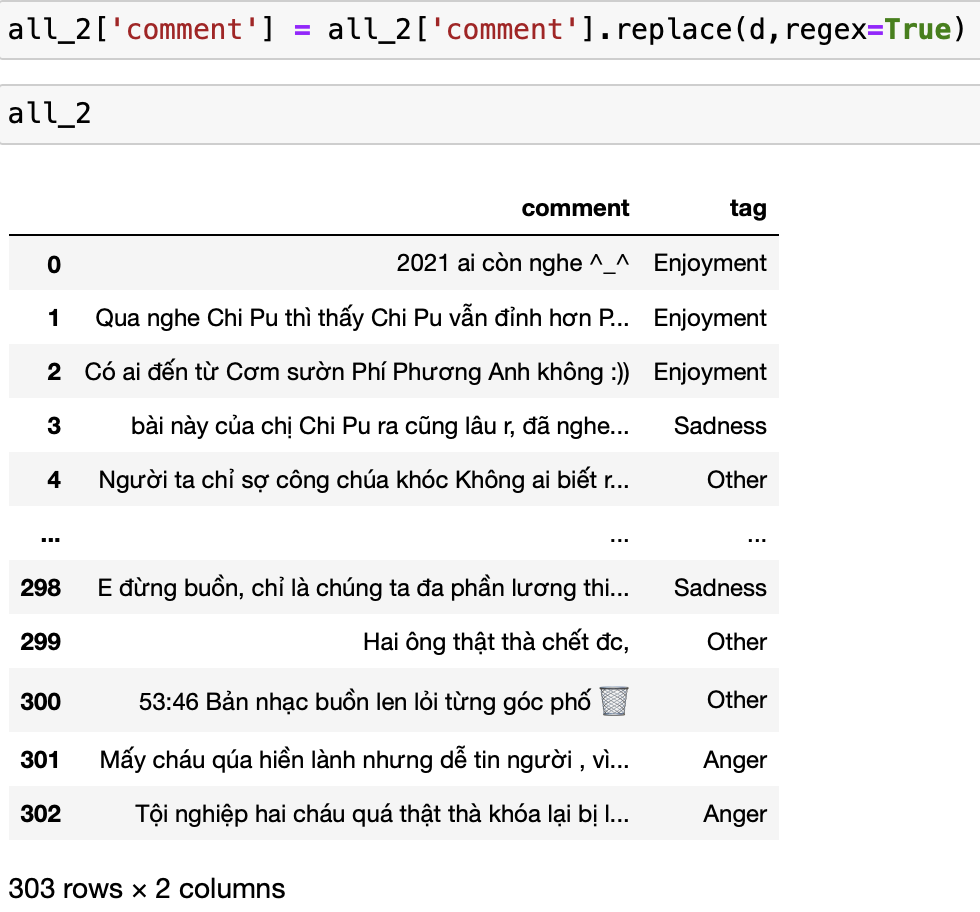
* Đọc 3 file Comments Youtube:



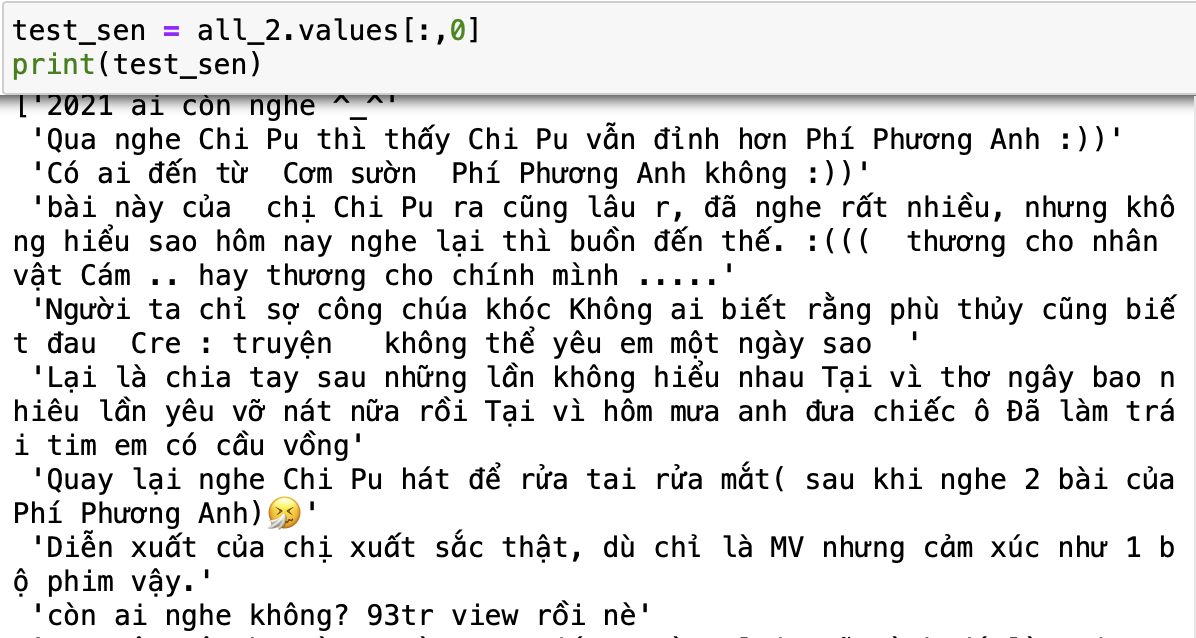
* Nối thành 1 file lớn sử dụng concat: (không sử dụng Merge bởi vì không nối được 2 file trở lên):



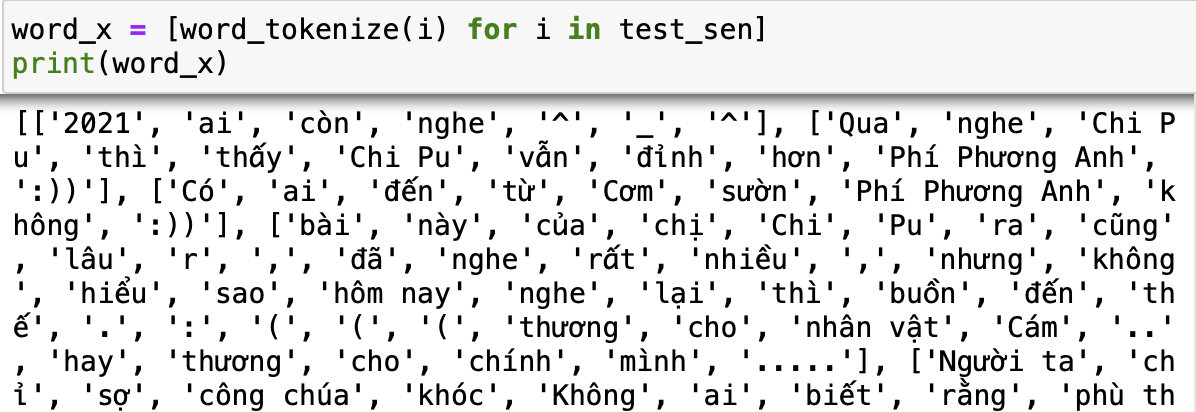
* Áp dụng sửa lỗi chính tả vào file:



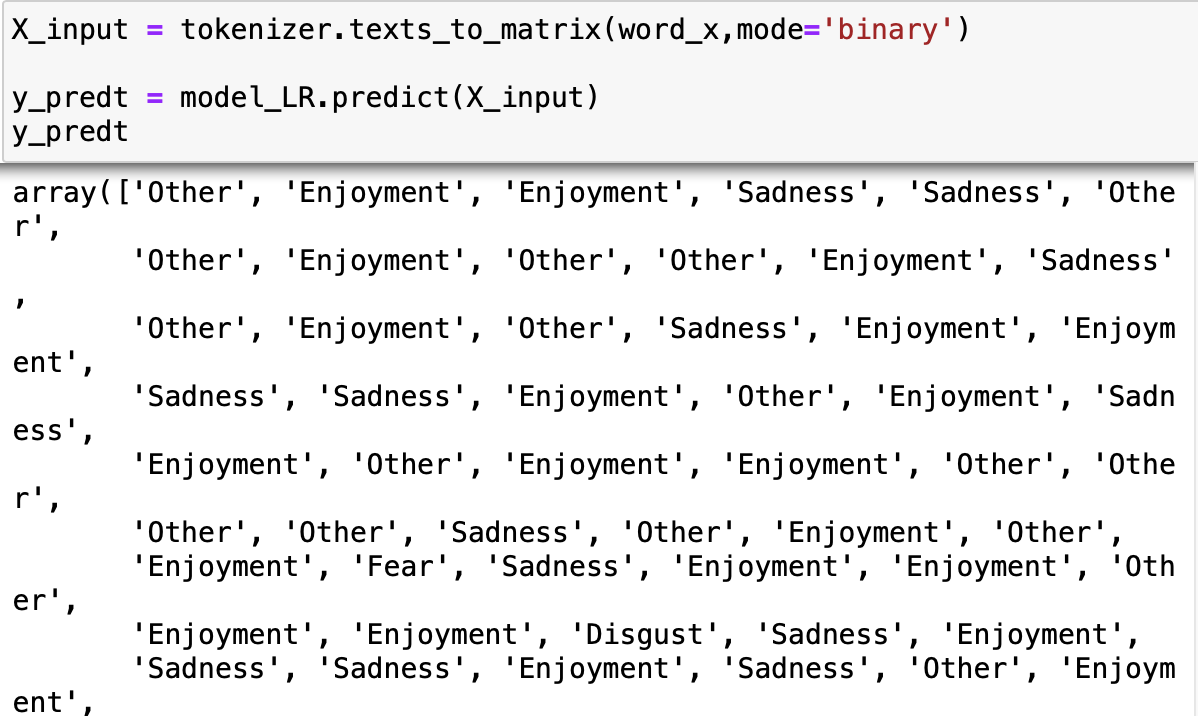
* Lấy dữ liệu các câu Comments:



* Áp dụng tách từ Tiếng Việt:



* Chuyển dữ liệu về ma trận mô hình Binary và dự đoán Cảm xúc:



* Tính độ chính xác của mô hình dự đoán so với dữ liệu được gắn tay:

