TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

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

**FINAL PROJECT**

*Giảng viên hướng dẫn*: **Thầy Lê Cung Tưởng**

*Người thực hiện*: **TRẦN TUẤN HUY – 51703103**

**NGUYỄN ANH CHÂU – 51703051**

Lớp **: 17050301**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

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

**FINAL PROJECT**

*Giảng viên hướng dẫn*: **Thầy Lê Cung Tưởng**

*Người thực hiện*: **TRẦN TUẤN HUY – 51703103**

**NGUYỄN ANH CHÂU – 51703051**

Lớp **: 17050301**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

# LỜI CẢM ƠN

Trong thời gian làm bài tập lớn, để tìm được tài liệu tham khảo theo đúng đề bài và những kỹ năng cần thiết để giúp cho việc hoàn thành bài tập lớn chúng em đã gặp khá nhiều khó khăn. Nhưng nhờ có sự gợi ý, hướng dẫn của thầy Lê Cung Tưởng - giảng viên lý thuyết môn Khai thác dữ liệu và khai phá tri thức về những tài liệu cần thiết để tham khảo cũng như kiến thức mà cô truyền đạt đã giúp em có những kiến thức, tài liệu, kỹ năng cần thiết để làm bài tập lớn. Dù sự giúp đỡ đó có ít hay nhiều, có gián tiếp hay trực tiếp thì chúng em vẫn vô cùng biết ơn cô.

Chúng em làm về đề tài Sentiment analysis - phân tích tâm lý của đối tượng, bài làm của chúng em còn nhiều hạn chế, thiếu sót. Do vậy, chúng em mong được cô góp ý để bài làm của chúng em hoàn thiện hơn.

Chúng em xin chân thành cảm ơn.

# BÀI TẬP LỚN ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là bài tập lớn của chúng tôi và được sự hướng dẫn của thầy Lê Cung Tưởng. Các nội dung tìm hiểu, kết quả trong đề tài này là trung thực. Những nội dung số liệu trong các sơ đồ là được viết dựa trên những gì chúng tôi tìm hiểu được.

Ngoài ra, trong bài tập lớn 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 bài tập lớn 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 15 tháng 4 năm 2021.*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trần Tuấn Huy*

*Nguyễn Anh Châu*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm 2021.

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

Tp. Hồ Chí Minh, ngày tháng năm 2021.

(kí và ghi họ tên)

# TÓM TẮT

Nội dung của bài tập lớn này là về đề tài bài tập lớn môn Khai thác dữ liệu và khai phá tri thức . Hiện nay, việc người dùng phản hồi ý kiến của mình về một vấn đề cụ thể nào đó thông qua ngôn ngữ tự nhiên có bản chất mơ hồ, khó đoán gây khó khăn cho việc xử lý bằng máy tính. Do đó để thuận tiện hơn cho việc phân tích phản hồi của người dùng, kĩ thuật Sentiment analysis đã ra đời. Trong bài tập lớn này, chúng em sẽ tìm hiểu về kỹ thuật này, cũng như hiện thực, so sánh kết quả của nó qua một số comment dataset từ người xem mà chúng em crawl được từ Youtube.

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc72077333)

[BÀI TẬP LỚN ĐƯỢC HOÀN THÀNH ii](#_Toc72077334)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc72077335)

[TÓM TẮT iv](#_Toc72077336)

[MỤC LỤC 1](#_Toc72077337)

[CHƯƠNG 1 – GIỚI THIỆU CHUNG 2](#_Toc72077338)

[1.1 Sentiment Analysis là gì? 2](#_Toc72077339)

[1.2 Phân tích tình cảm chi tiết 2](#_Toc72077339)

[1.3 Phân tích từng câu từng phần 2](#_Toc72077339)

[1.4 Tóm gọn 3](#_Toc72077339)

[CHƯƠNG 2 – GIẢI QUYẾT YÊU CẦU BÀI TOÁN 3](#_Toc72077341)

[2.1 Tổng quan về yêu cầu đồ án: 3](#_Toc72077359)

[2.2 Giải quyết bài toán: 4](#_Toc72077360)

[THÔNG TIN THAM KHẢO 42](#_Toc72077341)

CHƯƠNG 1 – GIỚI THIỆU CHUNG

* 1. Sentiment Analysis là gì?

Phân tích cảm xúc là một loại nghiên cứu văn bản hay còn gọi là khai thác. Nó áp dụng kết hợp thống kê, xử lý ngôn ngữ tự nhiên (NLP) và học máy để xác định và trích xuất thông tin chủ quan từ các tệp văn bản, chẳng hạn như cảm xúc, suy nghĩ, nhận định hoặc đánh giá của người đánh giá về một chủ đề, sự kiện hoặc một công ty cụ thể và các hoạt động của nó như đã đề cập ở trên. Loại phân tích này còn được gọi là khai thác ý kiến (tập trung vào việc trích xuất) hoặc đánh giá tình cảm. Một số chuyên gia cũng sử dụng thuật ngữ phân loại và trích xuất tình cảm. Bất kể tên gọi nào, mục tiêu của phân tích tình cảm đều giống nhau: biết ý kiến của người dùng hoặc khán giả về đối tượng mục tiêu bằng cách phân tích một lượng lớn văn bản từ nhiều nguồn khác nhau.

1.2 Phân tích tình cảm chi tiết:

Loại phân tích này được thực hiện ở cấp độ tài liệu và câu. Trên thực tế, hầu hết các chuyên gia sử dụng nó để phân tích các câu hơn là toàn bộ tài liệu. SA chi tiết thô đòi hỏi hai nhiệm vụ nhất quán: phân loại chủ quan và phát hiện và phân loại theo cảm tính.

* **Subjectivity** **classification**
* **Sentiment detection and classification**

Đôi khi mọi người chia sẻ quan điểm của họ mà không có cảm xúc. Ví dụ, tác giả của câu tôi nghĩ mọi người đều xứng đáng có cơ hội thứ hai bày tỏ ý kiến chủ quan của mình. Tuy nhiên, thật khó để hiểu chính xác cảm nhận của người viết về mọi người. Vì vậy, câu không thể hiện tình cảm và trung lập. Những câu trung lập - những câu thiếu tình cảm - thuộc loại độc lập không nên được coi là câu xen kẽ.

1.3 Phân tích từng câu từng phần

Bạn áp dụng phân tích chi tiết ở cấp độ câu phụ và nó có nghĩa là để xác định mục tiêu (chủ đề) của một tình cảm. Một câu được chia thành các cụm từ hoặc mệnh đề, và mỗi phần được phân tích theo mối liên hệ với những phần khác. Nói một cách đơn giản, bạn có thể xác định ai nói về một sản phẩm và chính xác thì một người nói về điều gì trong phản hồi của họ. Ngoài ra, nó giúp hiểu lý do tại sao một nhà văn đánh giá nó theo một cách nhất định.

Ví dụ, phân tích chi tiết rất hữu ích để xử lý các biểu thức so sánh (ví dụ: Samsung tốt hơn iPhone) hoặc các bài đăng ngắn trên mạng xã hội.

Nó không chỉ cho phép bạn hiểu cách mọi người đánh giá sản phẩm hoặc dịch vụ của bạn mà còn xác định tính năng hoặc khía cạnh mà họ thảo luận: Bàn di chuột trên máy tính xách tay của tôi đã ngừng hoạt động sau 4 tháng sử dụng. Bằng cách này, bạn biết chính xác những gì phải được cải thiện hoặc xem xét lại.

Khả năng xác định cường độ tình cảm là một lợi thế khác của phân tích chi tiết. Ngoài ba điểm tình cảm (tiêu cực, trung tính và tích cực), bạn có thể sử dụng các danh mục rất tích cực và rất tiêu cực.

1.4 Tóm gọn:

Sentiment analysis là một chủ đề thách thức trong Machine Learning. Người nói thể hiện cảm nhận của mình thông qua ngôn ngữ tự nhiên có bản chất mơ hồ đã gây không ít khó khăn để làm cho máy tính có thể hiểu được. Chưa kể, họ sử dụng các cách chơi chữ, ẩn ý hay các kí hiệu thể hiện cảm xúc như.

Vì vậy ở bài tập lớn này, chúng em sẽ tìm hiểu và hiện thực hóa kỹ thuật này.

CHƯƠNG 2 – GIẢI QUYẾT YÊU CẦU BÀI TOÁN



*2.1 Tổng quan về yêu cầu đồ án:*

Phần 1: Chọn 3 video trên Youtube**,** crawl 100 bình luận **ngẫu nhiên** từ mỗi video và gán nhãn:

* Disgust (chán ghét)
* Enjoyment(thích thú)
* Anger (giận dữ),
* Surprise (ngạc nhiên)
* Sadness (buồn bã)
* Fear (sợ hãi)
* Other (khác).

Phần 2: **S**ử dụng tập dữ liệu tiếng Việt để training cho model. Có thể dùng bất kì kỹ thuật nào và model nào.

Báo cáo kết quả dự đoán của model (Accuracy, F1-score...) bằng test set.

Phần 3: Áp dụng model ở phần 2 lên 3 dataset ở phần 1, báo cáo kết quả dự đoán của model model (Accuracy, F1-score...).

*2.2 Giải quyết bài toán:*

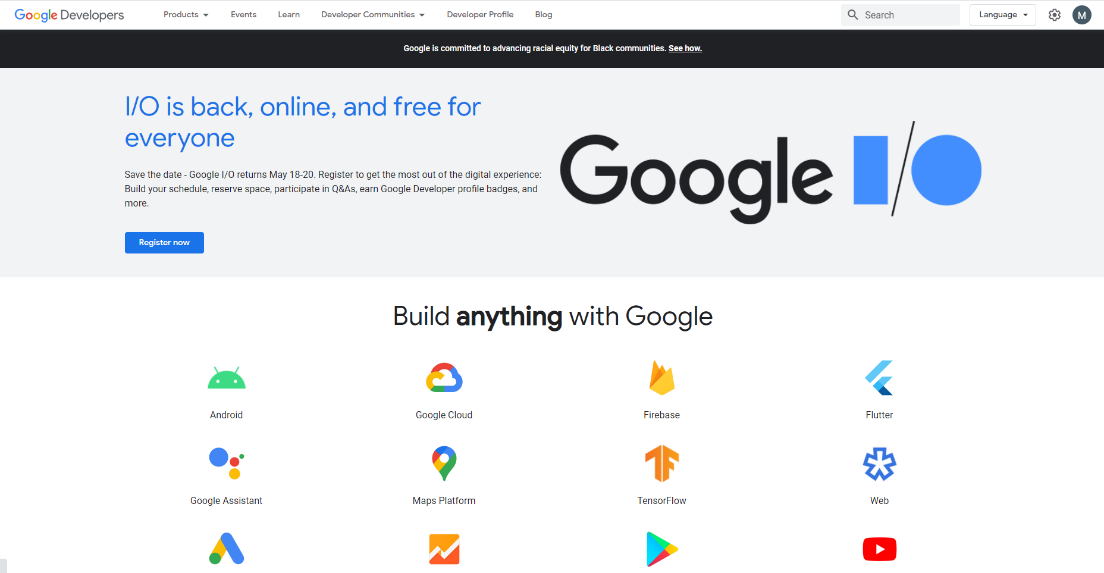
**Phần 1: Crawl bình luận từ video youtube**

**API là gì?** API (**Application Programming Interface)** - tức giao diện lập trình ứng dụng là phương tiện cho nhiều ứng dụng tương tác với nhau, tạo ra tương tác giữa người dùng với ứng dụng hiệu quả và tiện lợi hơn. Với API, các lập trình viên có thể tiếp cận, truy xuất dữ liệu từ máy chủ thể hiện chúng trên ứng dụng phần mềm hoặc website của mình một cách dễ dàng hơn.

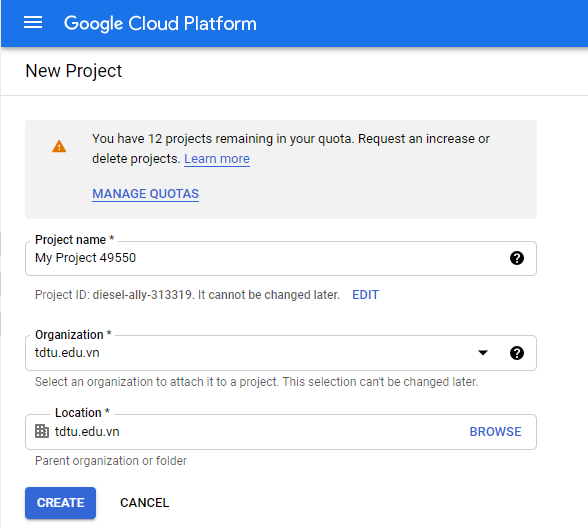
**Youtube API:**API **YouTube** giúp phá vỡ các rào cản giữa các hệ thống và đơn giản hóa các quy trình bằng cách nâng cao hiệu quả**,** có thể được sử dụng cho nhiều mục đích khác nhau bao gồm tìm kiếm nội dung, kiểm tra các yêu cầu được phép và trái phép, cũng như các tác vụ đơn giản khác như quản lý danh sách phát của bạn, tải video lên và cập nhật cài đặt kênh.

**API KEY:** Để thực hiện được phần 1 của bài toán, ta cần API key để sử dụng API của youtube:

1. Đăng nhập hoặc tạo tài khoản trên https://developers.google.com /.

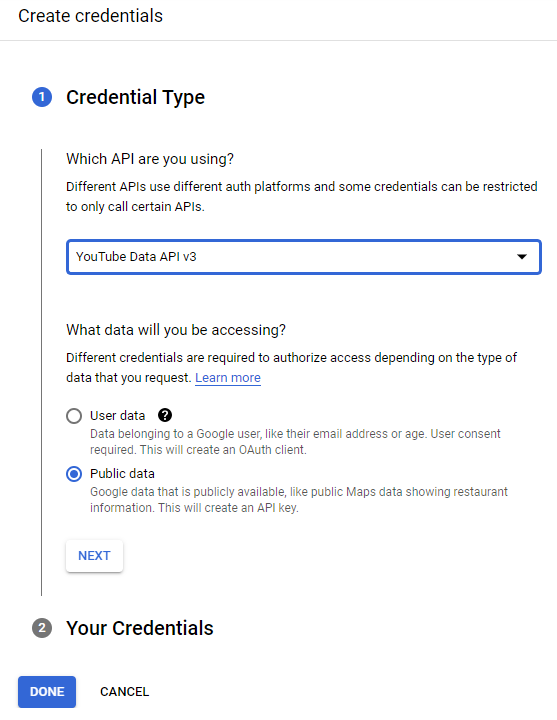


Vào <https://console.developers.google.com/>, tạo project mới

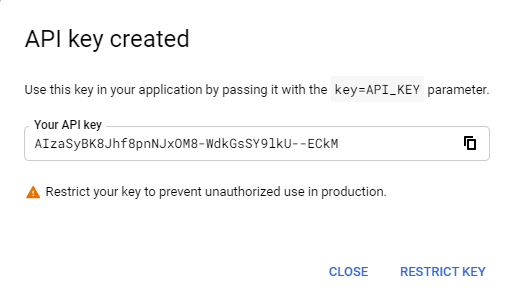


3. Sau khi tạo project, tìm YouTube Data API v3, chọn enable.

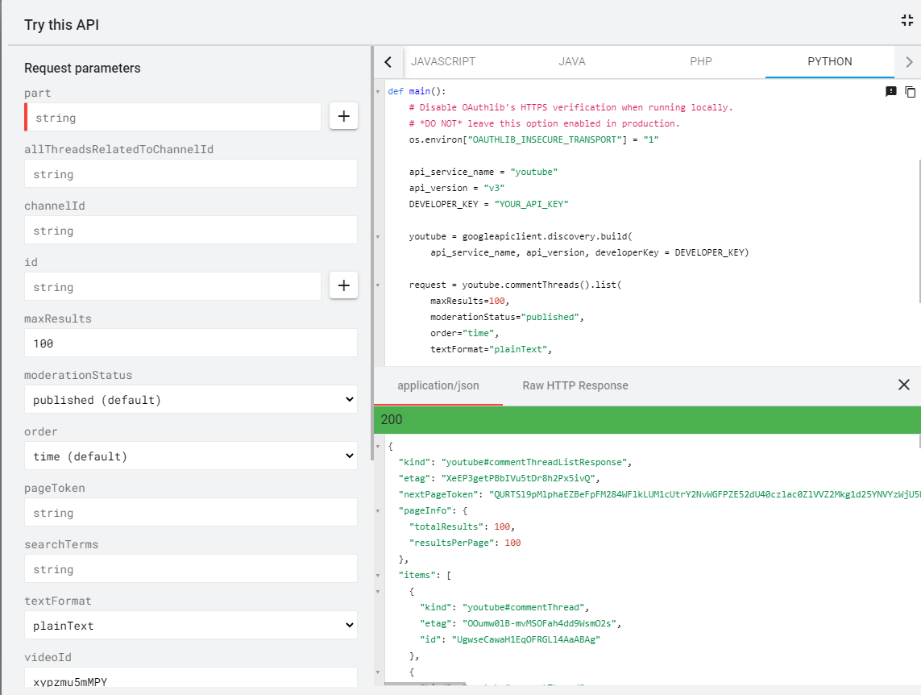
4. Chọn tab Credentials, Youtube Data API v3, check vào public data, chọn Done.



5. Click vào Create Credentials và chọn API Key. Hệ thống sẽ tự động tạo API key.

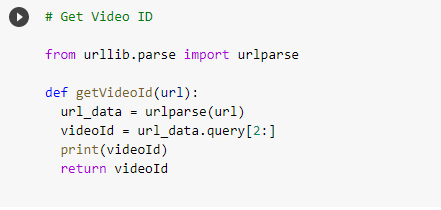


***Crawl bình luận từ Youtube API***



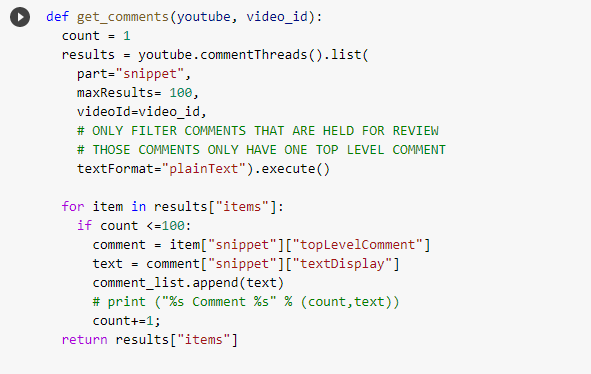
Để crawl được bình luận từ 1 video, chúng ta phải biết được videoID của video đó.

Hàm xử lý link để lấy video ID:

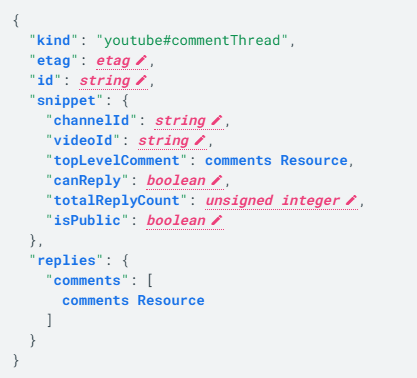
  


ID của một video cụ thể.

Hàm crawl 100 bình luận:



Hàm list của commentThreads để trả về list các items commentThreads resource có cấu trúc:

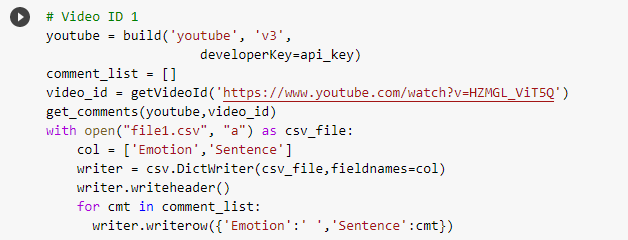


topLevelComment sẽ có dữ liệu là comments Resource – những thuộc tính của một bình luận như sau:

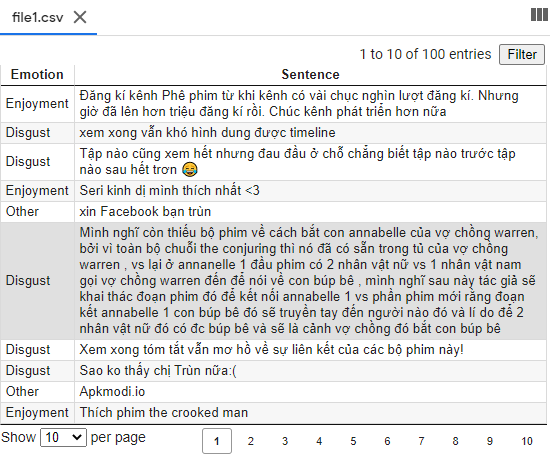


Ta lấy được bình luận thông qua **topLevelComment** của commentThreads resource và **textDisplay** của comments Resource bằng cách: Duyệt từng phần tử trả về của hàm **list trong** commentThreads sau đó thông qua nhánh ‘snippet’ ta có được các thuộc tính của một Comment, từ đó sẽ lấy ra được bình luận của người xem thông qua thuộc tính ‘textDisplay’ từ nhánh ‘snippet’ của comments Resource.

Ghi ra file csv:



Gán nhãn dữ liệu:

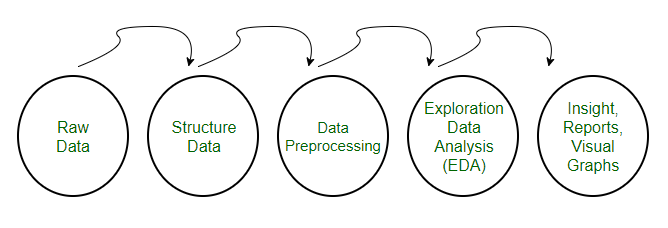


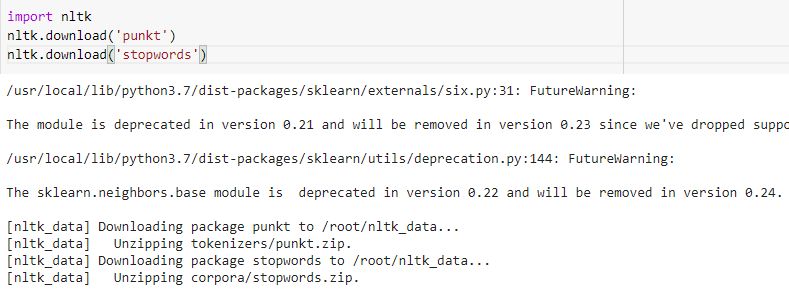
**PHẦN 2: Training Model từ dataset có sẵn**

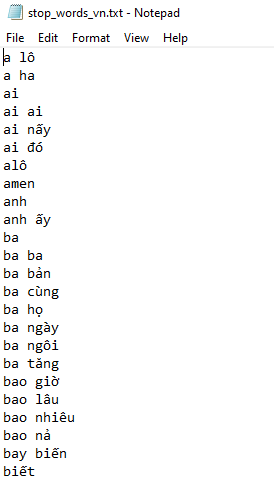
Tiền xử lý dữ liệu – Data preprocessing

Tiền xử lý được áp dụng cho dữ liệu trước khi đưa nó vào thuật toán.

Tiền xử lý dữ để chuyển đổi dữ liệu thô thành một tập dữ liệu sạch.

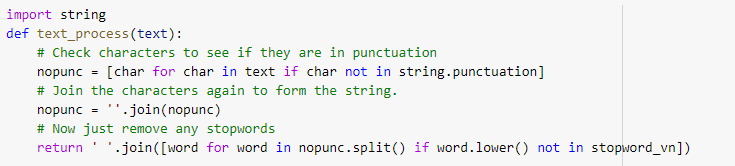
Download thư viện nltk cần thiết cho việc tiền xử lý dữ liệu



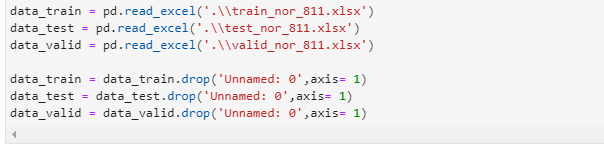
File stop\_words\_vn.

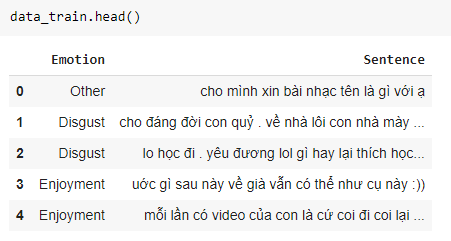
Tạo hàm create\_stopwordslist để chứa các stopword có trong file:

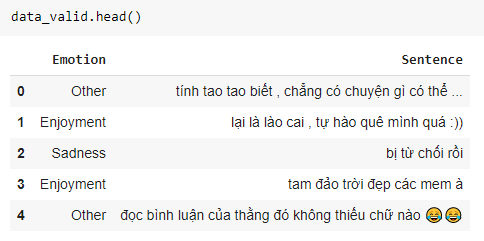


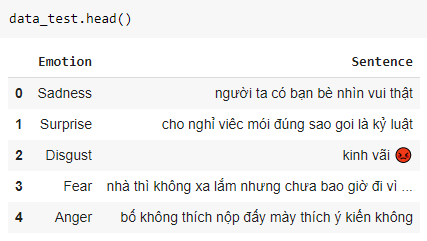
Tạo hàm xử lý dữ liệu để loại bỏ stopwords và các dấu câu

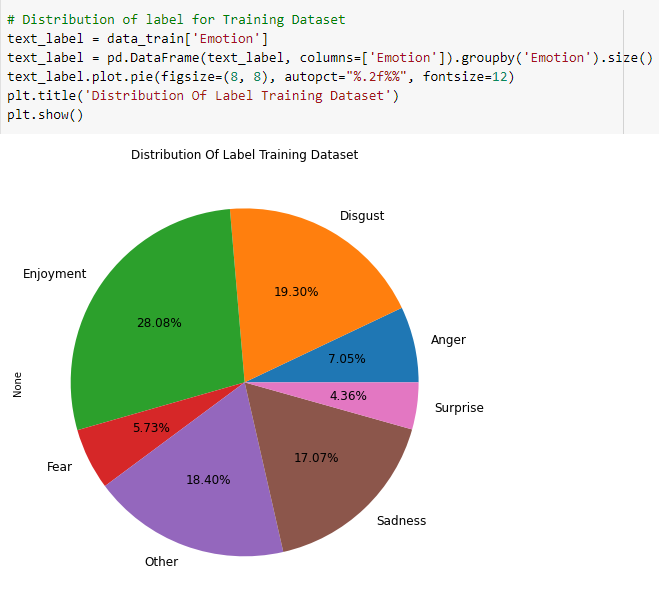
Sau đó ta sẽ đọc các file dữ liệu được cung cấp đồng thời xóa những cột không xác định







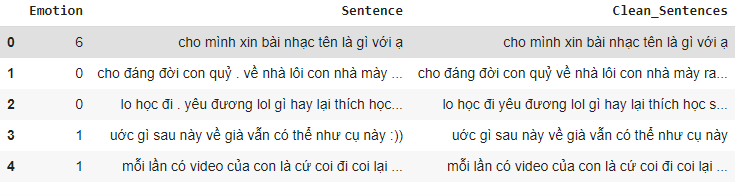




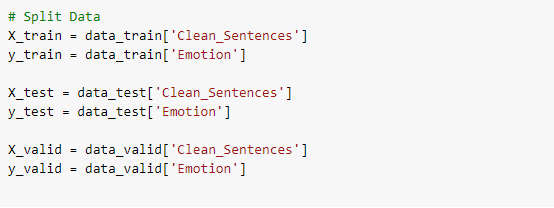
Ta dùng hàm map để chuyển nhãn thành số.

Disgust: 0, Enjoyment: 1, Anger: 2, Surprise: 3, Sadness: 4, Fear: 5, Other: 6

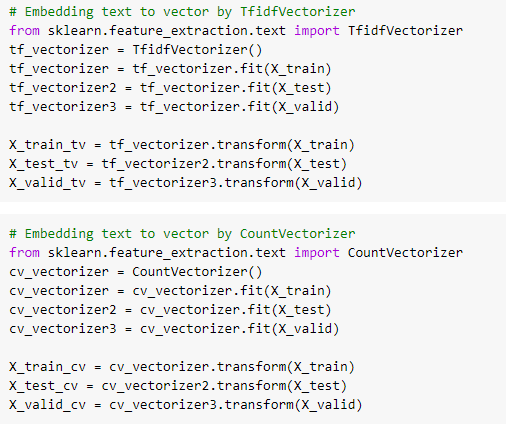
Dữ liệu có được sau khi tiền xử lý:



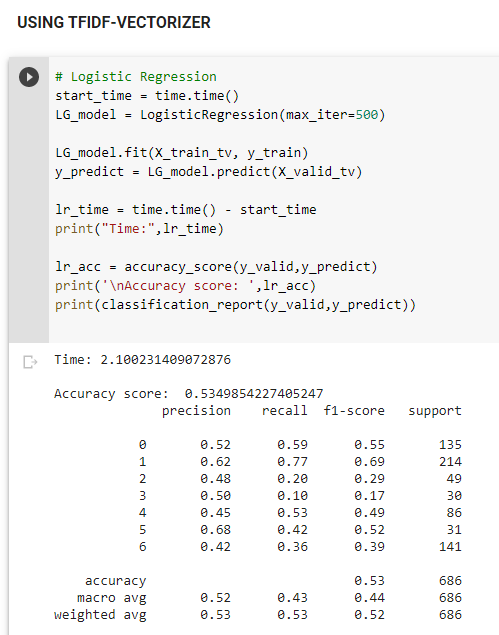
Ta chia dữ liệu ra để train, test, chuyển các dữ liệu về dạng các vector.



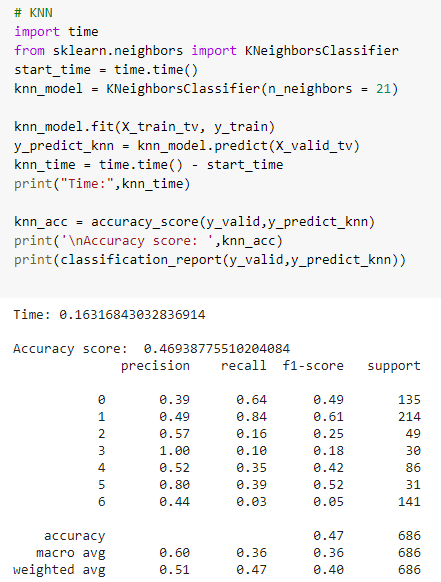
Ta sử dụng 2 loại vectors khác nhau: TFIDF-Vectorizer và CountVectorizer.



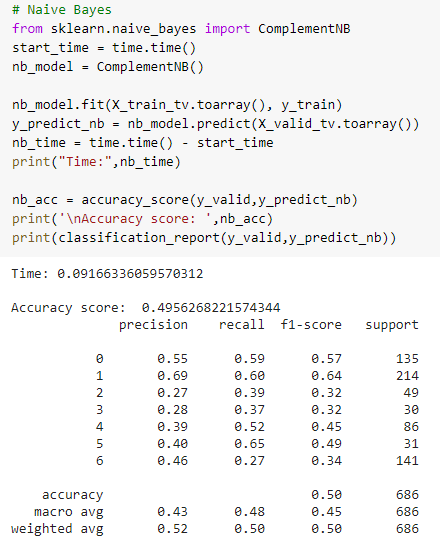
Đưa dữ liệu đã được xử lý vào model để training:



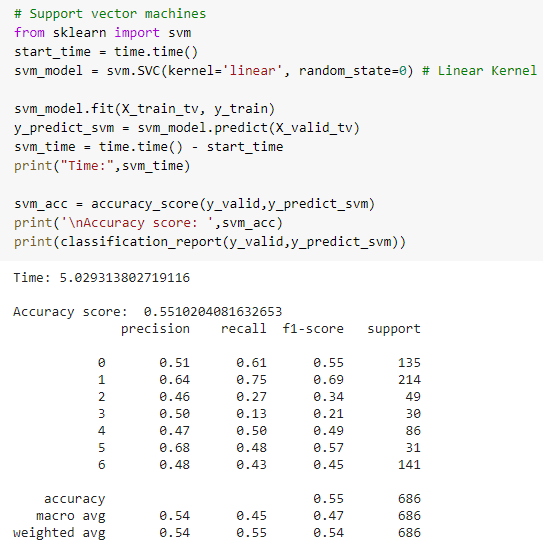
Logistic Regression sử dụng Tfidf-vectorizer là khoảng 53%.



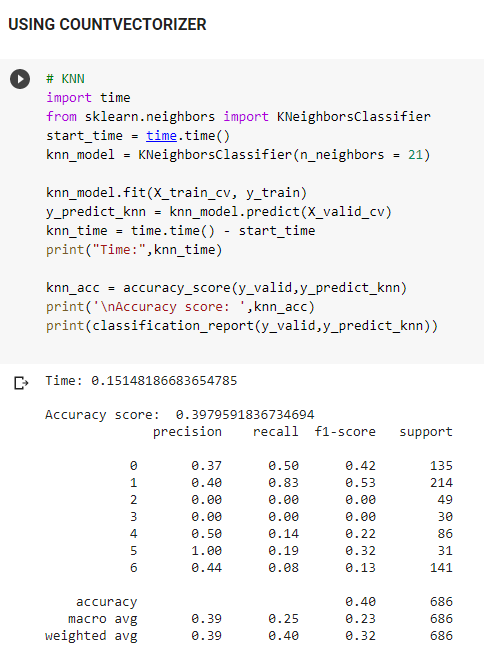
Model KNN là khoảng 47%



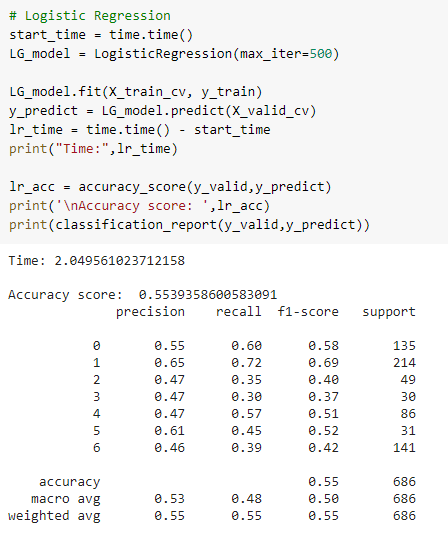
Naive Bayes là khoảng 49.5%



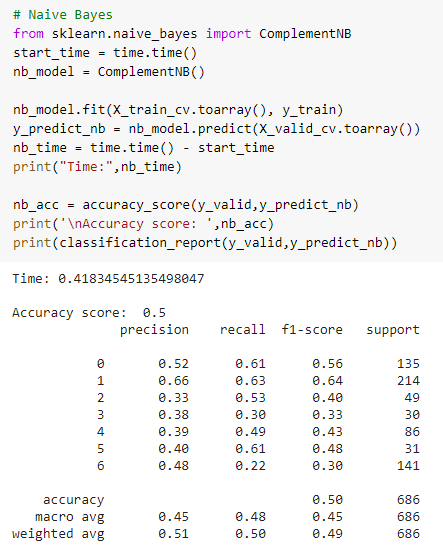
SVM là khoảng 55.1%



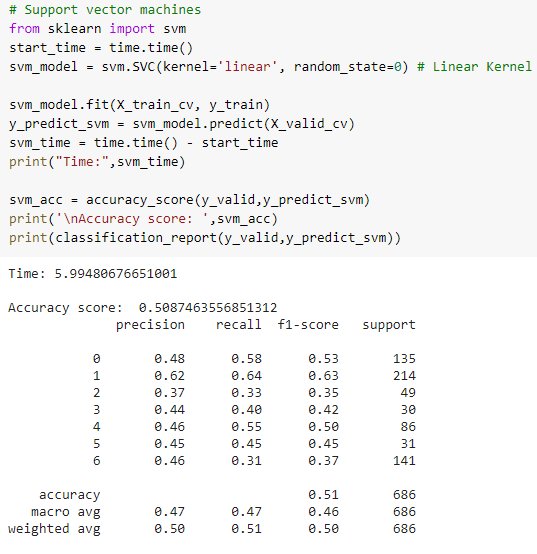
KNN sử dụng Count Vectorizer là khoảng 39.8%.



Logistic Regression sử dụng Count Vectorizer là khoảng 55.4%.

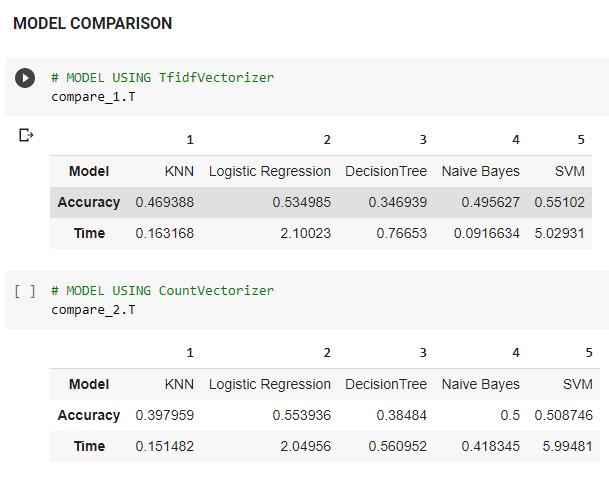


Naive Bayes dùng Count Vectorizer là 50%.



SVM dùng Count Vectorizer là 51%.

So sánh các kết quả dự đoán của các model khi sử dụng 2 loại vector khác nhau để training:



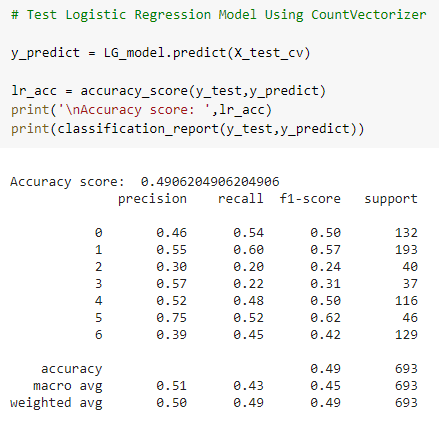
2 model có độ chính xác cao nhất là:

Logistic Regression dùng Count Vectorizer.

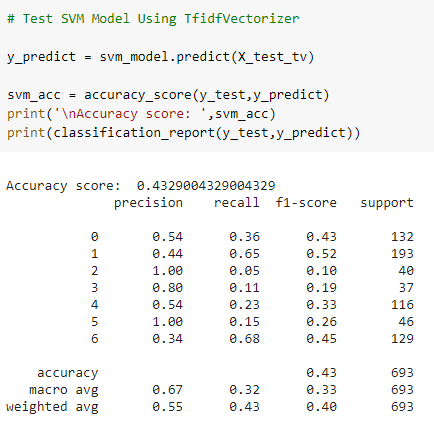
SVM dùng Tfidf vectorizer.

Chọn 2 model trên để sử dụng.

Dùng test data:



Logistic Regression – CountVectorizer: 49% cho 693 dữ liệu bình luận.



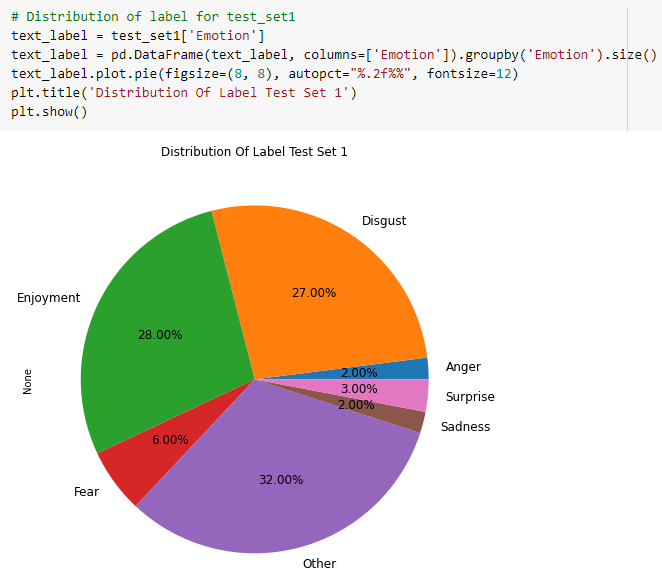
SVM - Tfidf Vectorizer: 43% cho 693 dữ liệu bình luận

**PHẦN 3: Áp Dụng Model Đã Training Cho Dataset Ở Task 1**

Tương tự như ở Task 2, trước khi áp dụng model đã được training cho dữ liệu ở Task 1. Chúng ta phải qua các bước tiền xử lý dữ liệu.

Đọc File bình luận từ clip thứ 1:



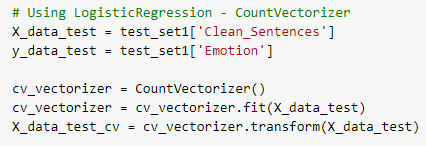
Thống k**ê**:

**C**huyển các nhãn thành số, xóa bỏ các stopwords và dấu câu các comments:

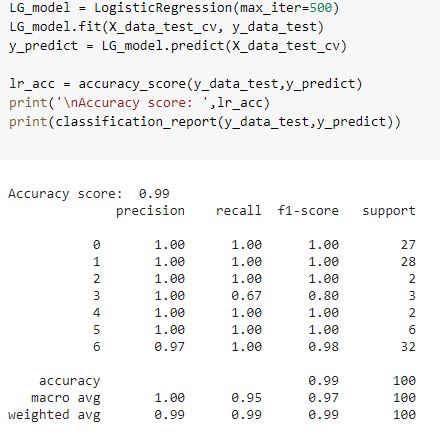


Ta chia dữ liệu thành nhãn và bình luận.

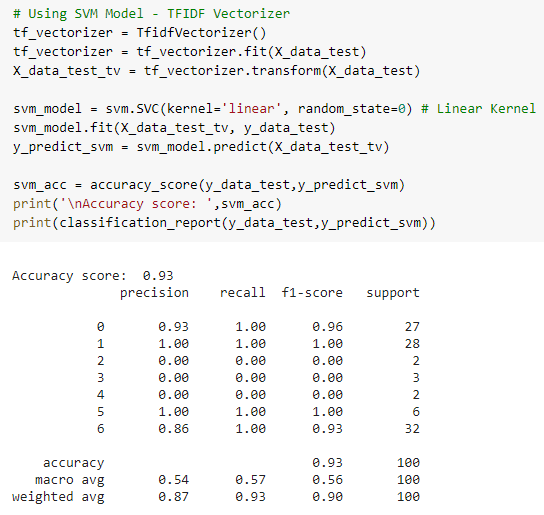
Chuyển các bình luận về dạng các vector bằng CountVectorizer.

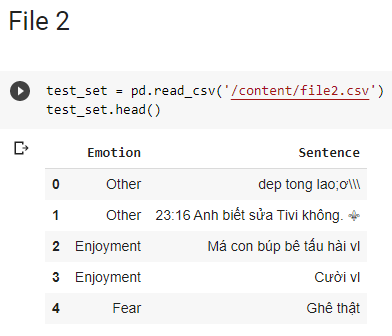


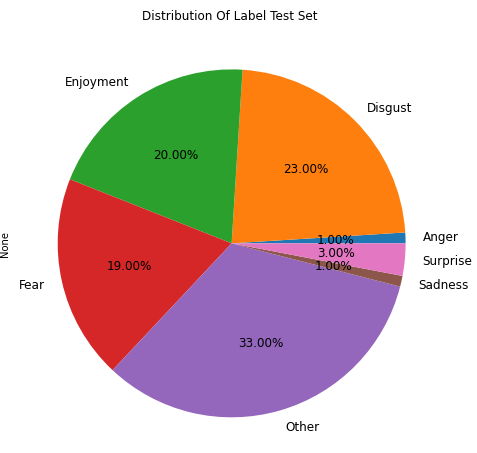
**D**ùng Logistic Regression để test và báo cáo kết quả:

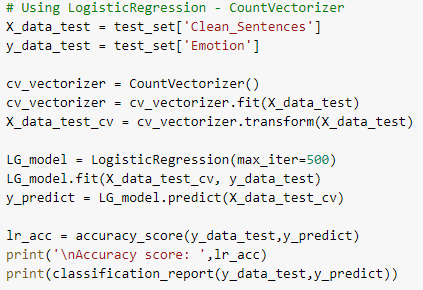


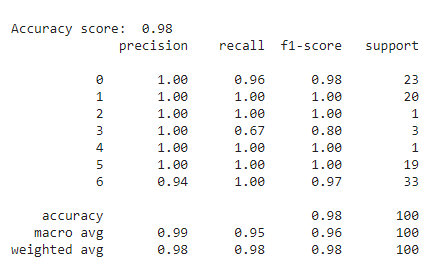
Tương tự cho model thứ 2 (SVM – TfidfVectorizer):

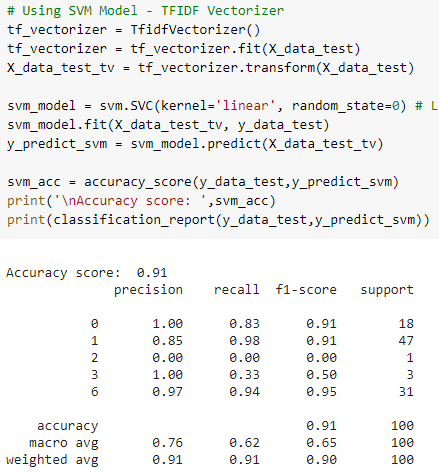


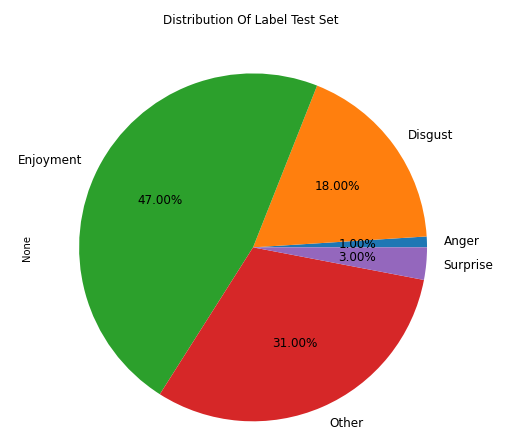
Tương tự ở File thứ 2, chúng ta sẽ crawl comments của người xem về video reviews phim ma Chucky:

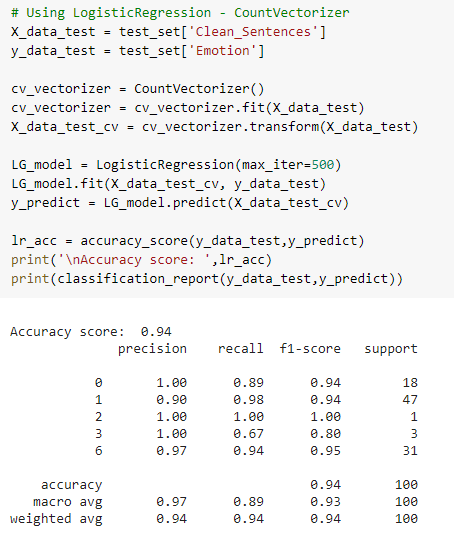


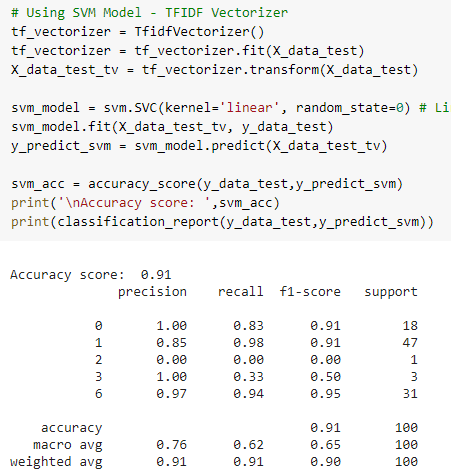




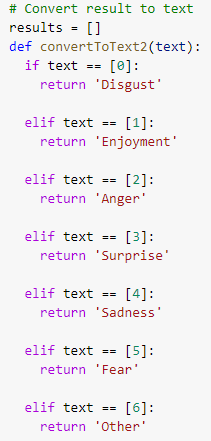


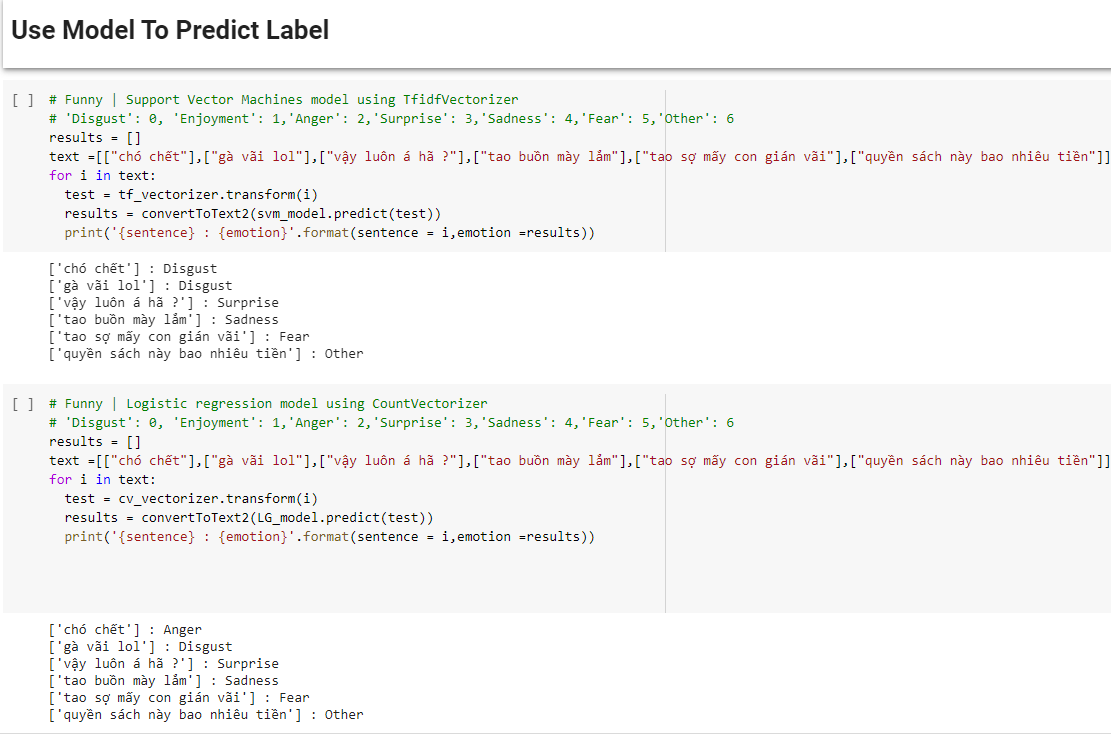
 





Ta có thể dùng 2 model trên để dự đoán một số bình luận

Convert thành chữ:

Kết quả:

Logistic Regression có phần chính xác hơn SVM!

Thông Tin Tham Khảo

* <https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/>
* <https://mona.media/api-la-gi/>
* <https://stackoverflow.com/>
* https://topdev.vn/