THU THẬP VÀ XỬ LÝ DỮ LIỆU CÁC BÌNH LUẬN YOUTUBE

1.1 Thu thập dữ liệu bình luận youtube

Để thu thập dữ liệu bình luận youtube, ta sẽ dùng thư viện selenium bằng câu lệnh: pip install selenium.



Hình 1.1 Cài đặt thư viện selenium

Tiến hành tạo selenium webdriver để thực hiện việc kiểm thử trực tiếp bằng trình duyệt web. Sau khi tạo xong, ta tạo tên biến “link\_youtube” chứa đường link youtube cần thu thập dữ liệu và dùng driver.get() để mở trình duyệt.



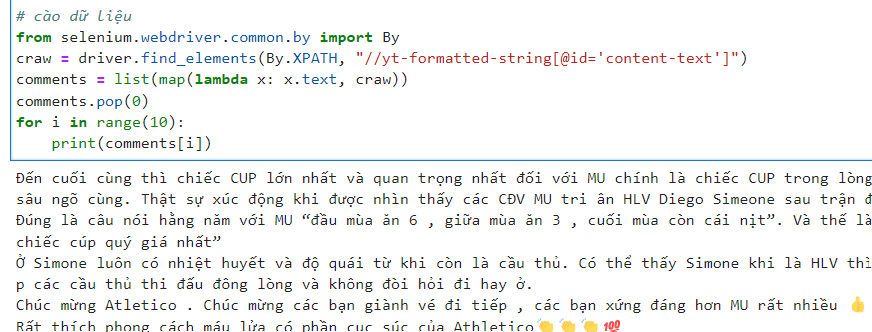
Hình 1.2 Tạo selenium webdriver

Do youtube sử dụng cơ chế cuộn trang tới đâu load dữ liệu tới đó, vì vậy ta cần phải tiến hành cuộn trang bằng cách sử dụng câu lệnh driver.execute\_script(), sau đó mới tiến hành bước thu thập dữ liệu.



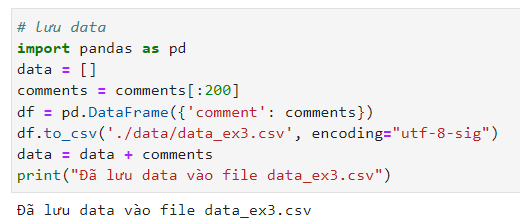
Hình 1.3 Cuộn trang

Tiến hành thu thập dữ liệu bằng lệnh driver.find\_elements() và lưu giá trị cào được vào biến “comments”



Hình 1.4 Thu thập dữ liệu

Cuối cùng, ta lưu dữ liệu vào file csv.



Hình 1.5 Lưu dữ liệu vào file csv

1.2 Tiền xử lý dữ liệu và dùng TF-IDF để trích xuất đặc trưng

1.2.1 Tiền xử lý dữ liệu

Để có thể sử dụng TF-IDF trích xuất đặc trưng, ta cần phải tiền xử lý dữ liệu trước. Đây là bước quan trọng để dữ liệu đầu vào đạt được chất lượng tốt nhất nhằm mục đích dữ liệu không chứa các kí tự đặc biệt, bị khuyết, từ không có ý nghĩa,...

Đầu tiên, ta cài đặt các thư viện cần thiết như underthesea, thư viện này giúp cho các từ riêng rẽ trong dữ liệu tiếng việt trở nên có ý nghĩa.



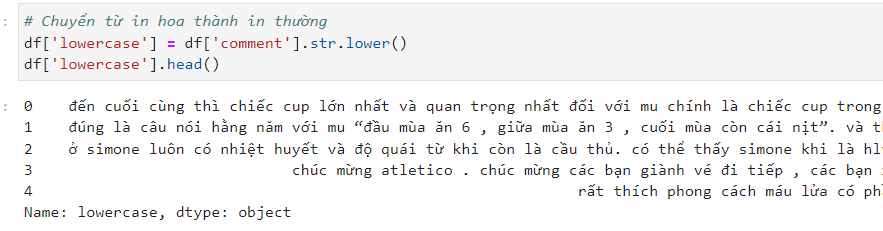
Hình 1.6 Cài đặt thư viện underthesea

Ta đọc dữ liệu file csv từ thu thập dữ liệu trước đó, lưu vào biến “df”



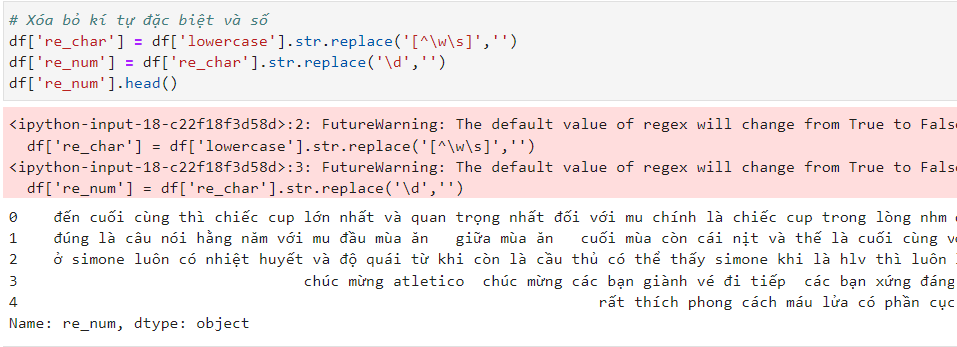
Hình 1.7 Đọc file csv

Bước đầu tiên trong tiền xử lý dữ liệu là chuyển các từ in hoa thành các từ in thường, điều này nhằm giúp các từ trong dữ liệu có độ thống nhất với nhau.



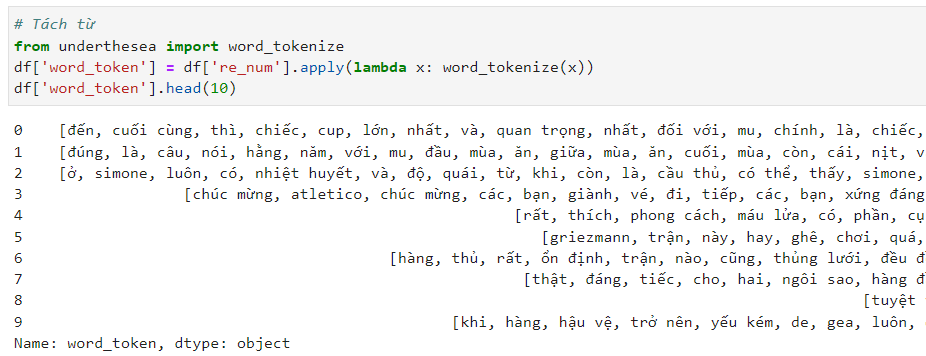
Hình 1.8 Chuyển từ in hoa thành in thường

Tiếp theo, ta xóa bỏ các kí tự đặc biệt như các icon, dấu câu và chữ số.



Hình 1.9 Xóa bỏ kí tự đặc biệt và số

Sử dụng thư viện underthesea để tiến hành tách từ trong dữ liệu.



Hình 1.10 Tách từ

Cuối cùng, sử dụng stop word để lọc các từ không phải là từ khóa, từ trọng tâm của dữ liệu. Dữ liệu trong file stopwords.txt được lấy từ:

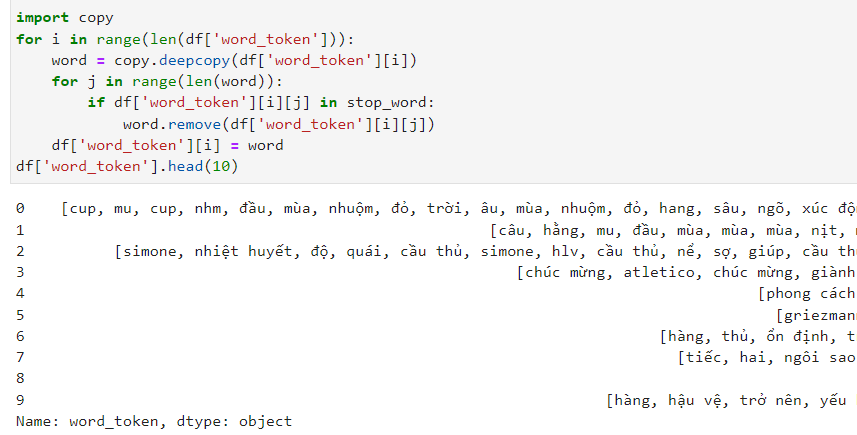
<https://github.com/stopwords/vietnamese-stopwords>

Để tiến hành stop word, ta phải đọc dữ liệu từ file stopwords.txt.



Hình 1.11 Đọc file stopwords.txt

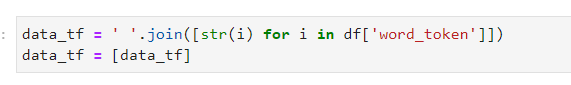
Ta sẽ copy giá trị của df[i] vào biến “word”, sau đó chạy vòng lặp trong biến “word”, nếu giá trị df[i][j] nằm trong stop word thì sẽ tiến hành remove giá trị đó. Kết quả sau khi chạy vòng lặp sẽ được lưu ngược lại về biến df[i].



Hình 1.12 Xử lý stop word

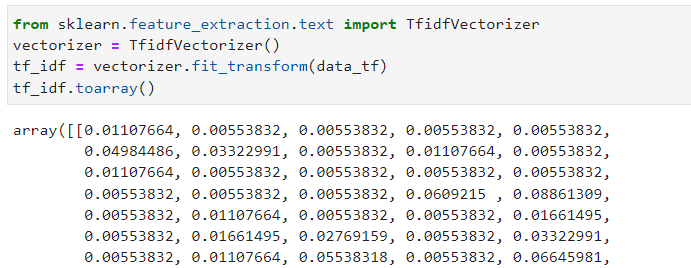
1.2.2 Trích xuất đặc trưng sử dụng TF-IDF

Đầu tiên, ta sẽ chuyển dữ liệu tiền xử lý trước đó thành dạng chuỗi (String).



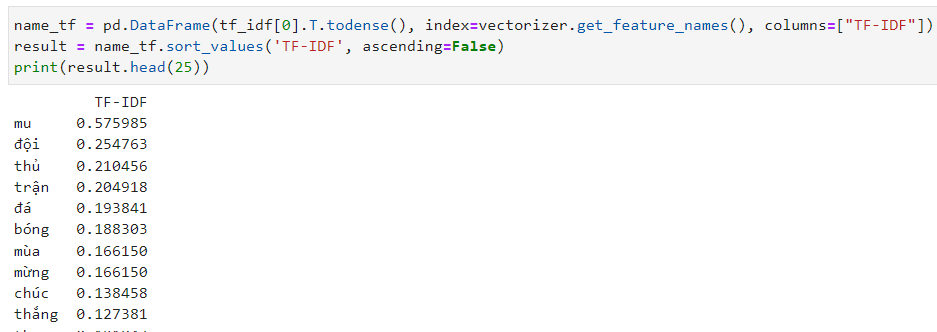
Hình 1.13 Chuyển dữ liệu thành dạng chuỗi

Để trích xuất đặc trưng sử dụng TF-IDF, ta sẽ dùng thư viện scikit-learn để hỗ trợ với class là sklearn.feature\_extraction.text.TfidfVectorizer, class này được dùng để chuyển dữ liệu thô thành dữ liệu ma trận sử dụng TF-IDF.



Hình 1.14 Trích xuất đặc trưng sử dụng TF-IDF

Từ ma trận có được như trên, ta sẽ tiến hành tạo danh sách tần suất xuất hiện của các từ trong từng dòng dữ liệu.



Hình 1.15 Tạo danh sách tần suất xuất hiện các từ trong từng dòng dữ liệu

Từ danh sách, ta thấy rằng từ “mu” có tần suất xuất hiện cao nhất với 0.57. Điều này đúng với trang web cào được do trang web nói về highlight của trận Man United – A. Madrid. Từ “mu” đây là nhắc về đội Man United, nếu ta xem các dòng bình luận thì cũng có thể thấy rằng đa số các dòng bình luận đều bình luận đến đội bóng này. Ngoài ra, những từ có tần suất cao hơn 0.16 cũng liên quan về chủ đề bóng đá, đây là những từ đặc trưng trong dữ liệu để từ đó có thể làm tiền đề cho các bước xử lý dữ liệu sau này.