–TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP GIỮA KỲ**

**HỌC MÁY**

*Người thực hiện*: **Đặng Thọ Duy – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**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**



**BÀI TẬP GIỮA KỲ**

**HỌC MÁY**

*Người thực hiện*: **Đặng Thọ Duy – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**MỤC LỤC**

***Trang***

**Text Classification**

1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu

1.2 Sơ đồ tổng quát

1.3 Đặc trưng của mô hình

1.3.1. Tiền xử lý

1.3.2 Feature

1.3.3 Giải thuật

1.3.4 Cách đánh giá

1.4 Thực nghiệm

1.4.1 Công nghệ sử dụng

1.4.2 Các tham số của mô hình

1.4.3 Kết quả đạt được

1.4.4 Thảo luận

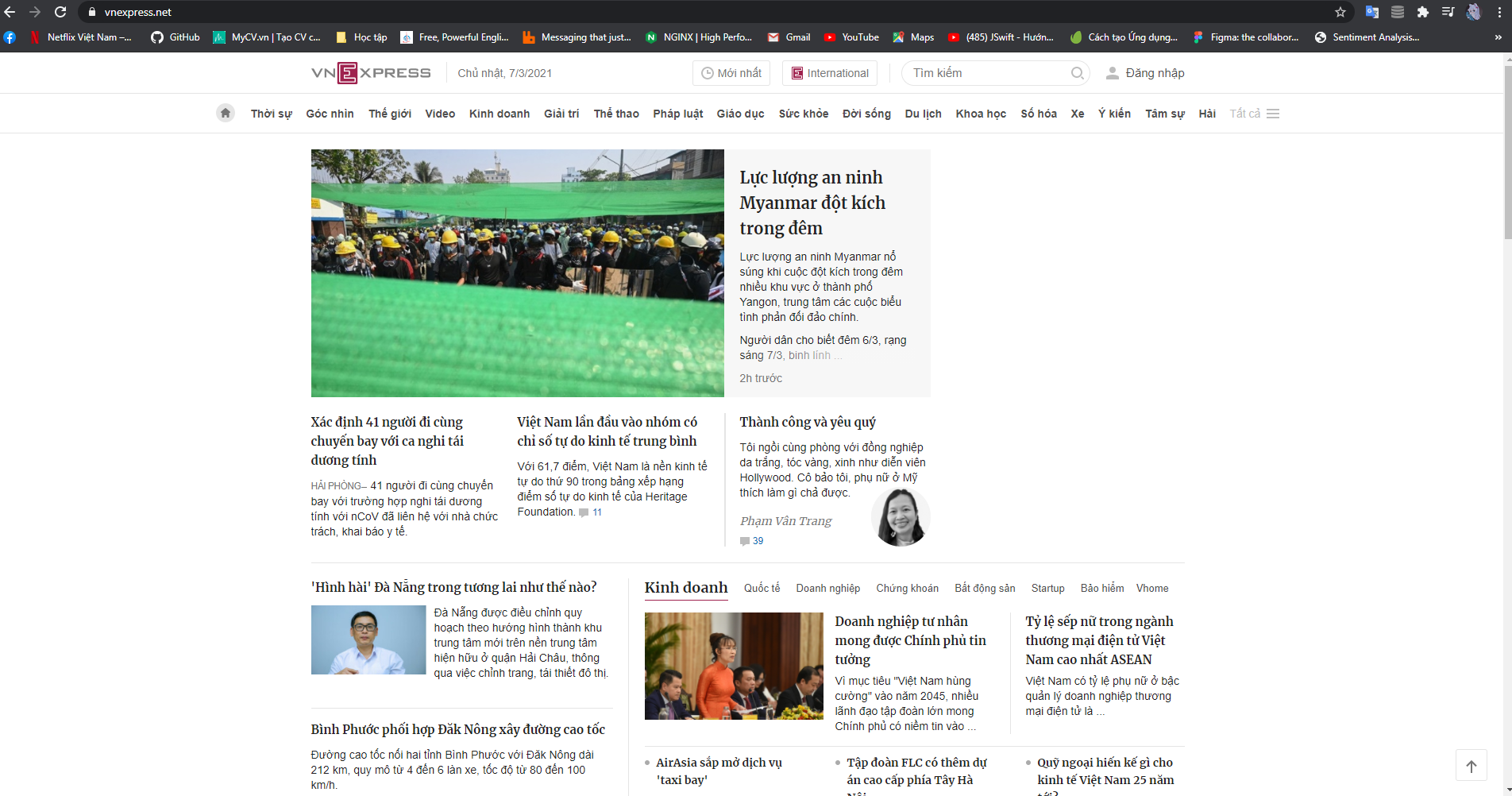
**Tự đánh giá**

**TEXT CLASSIFICATION**

**1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu**

Trình bày chi tiết đã xây dựng dữ liệu như thế nào?

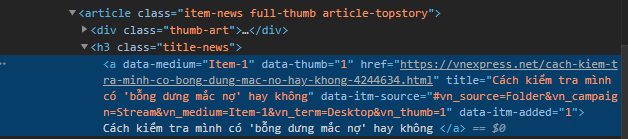
* Dữ liệu được crawl về từ trang báo online VNEXPRESS.NET



* Tương ứng với môi topic thì sẽ ứng với mộtk URL khác nhau, ở đây ta tiến hành chọn 5 topic.



* Bằng cách sử dụng thư viện request gửi đến các URL tương ứng với các topic. Ta sẽ mô phỏng được một người đang truy cập trang web đó.
* Tiếp đến ta dung thư viện BeautifulSoup (response.content) để lấy tất cả các nội dung trong trang và gán vào biến soup.
* Với mỗi topic như thế ta cần phải lấy Title và nội dung của từng bài báo, như vậy sau khi dung development tool ta thấy được với mỗi title của bài báo được lưu trữ trong thẻ h3 với class (title-news và title\_news) và đường link tới chi tiết bài báo được lưu ở thẻ con “<a>”.



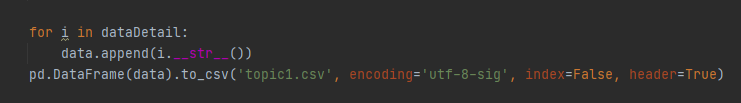
* Chỉ cần dung findAll với biến soup lúc nãy ta sẽ lấy tất cả các thẻ h3 với class tương ứng, và find từng thẻ con “a” bên trong thẻ cha “h3” để lấy đường link đến chi tiết từng bài báo.



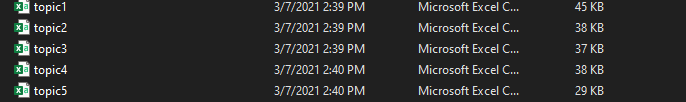
* Với từng chi tiết bài báo bên trong ta lặp lại như lúc nãy để lấy được Tiêu đề và Nội dung của bài báo.



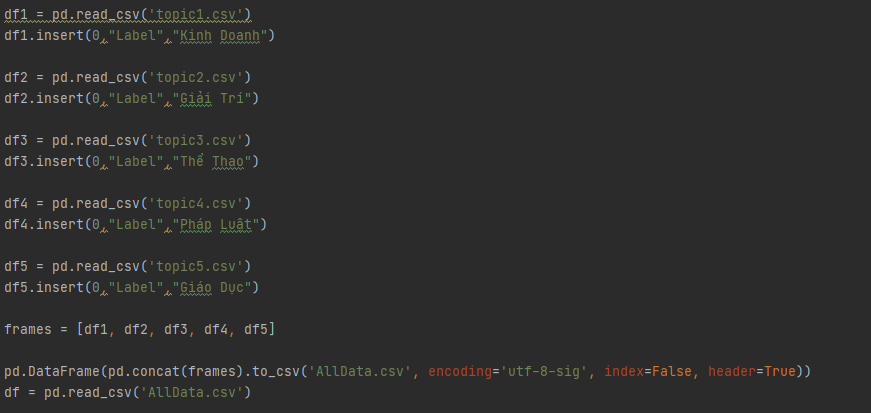
* Sau bước này là ta đã có chi tiết từng bài báo rồi. Công đoạn tiếp theo là chúng ta lưu trữ vào từng tệp với mỗi topic tương ứng.



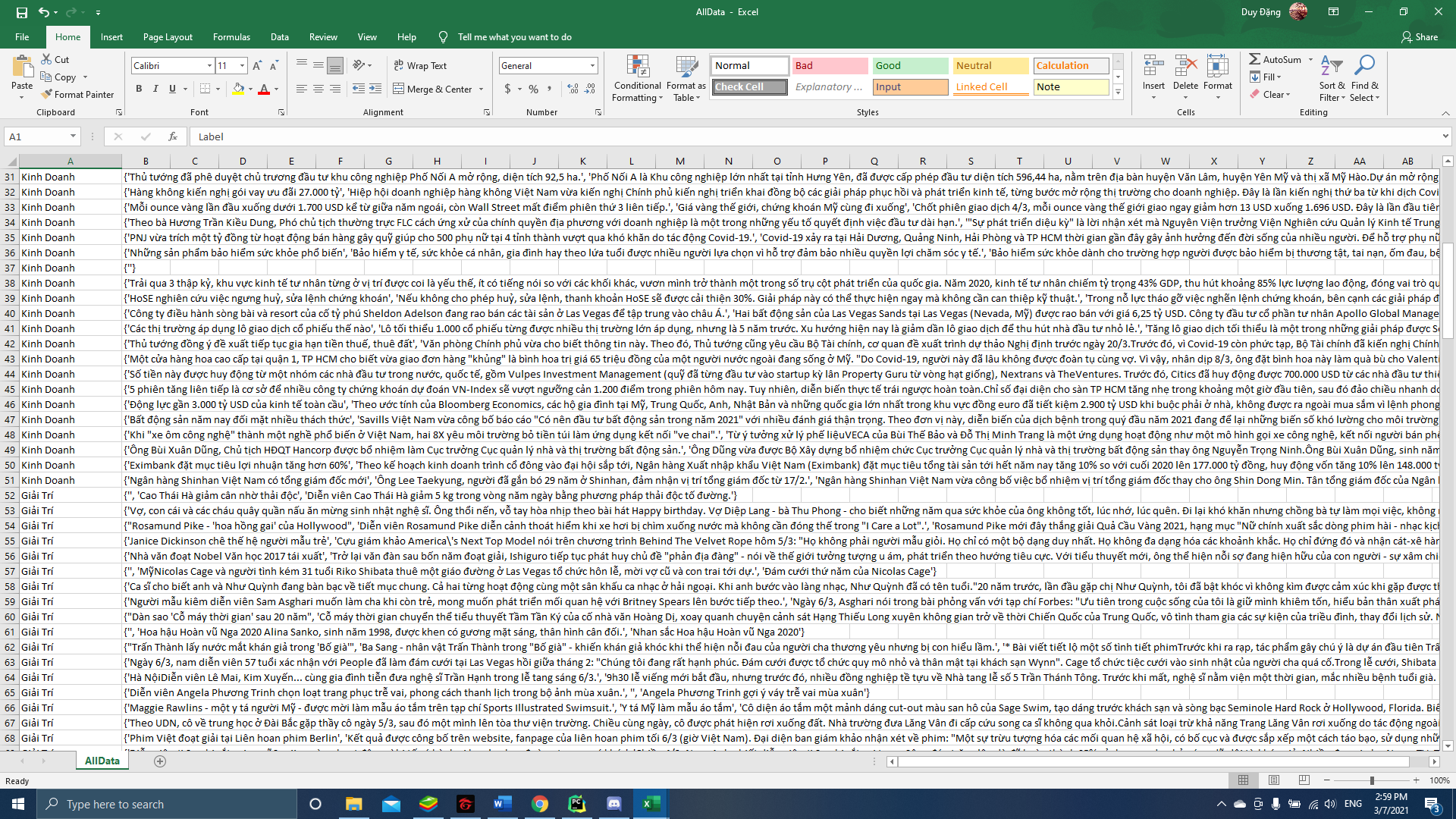
* Như thế này là ta đã xong 1 topic tương tự với các topic còn lại. Việc dung vòng for có thể hoàn toàn crawl nhiều topic một lần nhưng với số lượng ít thì mình coppy cho nhanh.
* Dữ liệu dung lượng bao nhiêu, bao nhiêu tập tin, bao nhiêu lớp
* Như vậy sau khi ta crawl dữ liệu về với mỗi tương ứng ta sẽ lưu vào các tập tin, ở đây mình có 5 topic thì tương ứng là 5 file.



* Nhưng đến đây chưa xong, để gọn hơn và gắn nhãn để train ta cần đọc các file này ra và nối chúng nó lại, rồi gán các nhãn vào.



* Như vậy sau khi gắn nhãn ta sẽ có một file chứa toàn bộ dữ liệu như thế này.



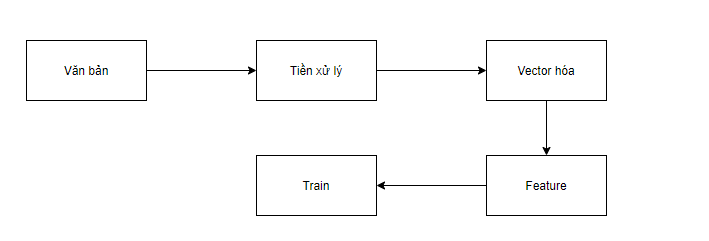
Phân tích dữ liệu (văn bản dài nhất, ngắn nhất, số từ, …)

Cách phân chia dữ liệu: Train, Test, Validation

* Phân chia dữ liệu: với mỗi bài báo sẽ gắn nhãn tướng ứng và bắt đầu với tỉ lệ 8:2

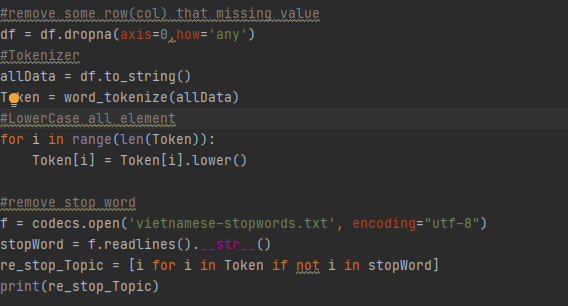
Phân tích các bộ dữ liệu trên

**1.2 Sơ đồ tổng quát**

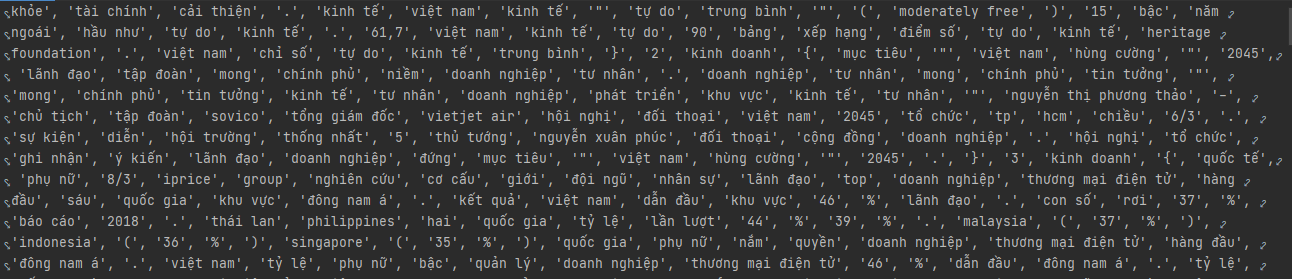


**1.3 Đặc trưng của mô hình**

1.3.1 Tiền xử lý



* Công đoạn tiền xử lý này bắt đầu với việc ta cần phải đọc lại file chứa toàn bộ dữ liệu lúc nãy bằng dataframe của thư viện pandas và dropna để xóa bỏ các dòng bị trống.
* Tiếp đến ta chuyển về dạng string và dung thư viện underthesea và tokenize chuỗi dữ liệu và lưu vào biến Token.
* Tiếp đến ta đọc từng từ và đưa về dạng chữ thường (không viết hoa).
* Cuối cùng là công đoạn loại bỏ stop word. Ở công đoạn này ta dung các stopword tiếng việt được liệt kê sẵn trong từ điển vietnammese-stopword (<https://github.com/stopwords/vietnamese-stopwords>). Tải bộ từ điển này về và đọc từng dòng rồi so sánh với Token lúc nãy, và từ nào giống nhau thì ta loại bỏ đi.
* Như vậy công đoạn tiền xử lý đã xong và bộ dữ liệu ta thu lại được như sau:

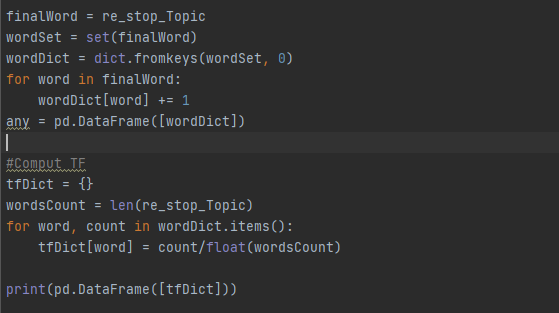


1.3.2 Feature

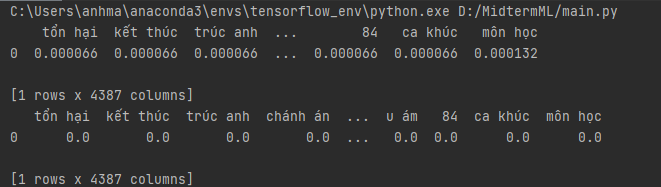
* TF – IDF (Term Frequency – Inverse Document Frequency) là thuật toán đánh giá tầm quan trọng của từ trong một văn bản. Giá trị càng cao thể hiện độ quan trọng càng cao và nó phụ thuộc vào số lần xuất hiện trong văn bản nhưng bù lại bởi tần xuất trong dữ liệu.
* TF: Term Frequency là số lần từ xuất hiện trong văn bản.



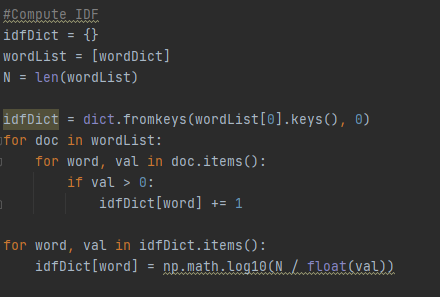
* F(t,d) -số lần xuất hiện từ t trong văn bản d.
* Max{f(w,d):**w∈d**} – số lần xuất hiện nhiều nhất của một từ bất kì trong văn bản.
* Với chuỗi token lúc nãy ta áp dụng công thức:



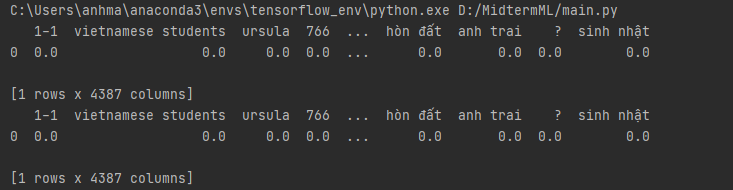
* Kết quả:



* IDF: Inverse Document Frequency( Tần suất nghịch đảo của văn bản), giúp đánh giá quan trọng của một từ.



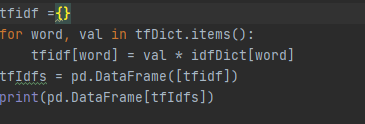
-Kết quả:



-TF-IDF:

Khi có cả TF và IDF ta dễ dàng tính được TF-IDF

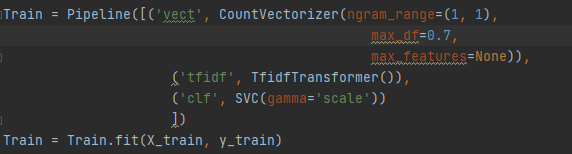




* Kết Quả:

1.3.3 Giải thuật

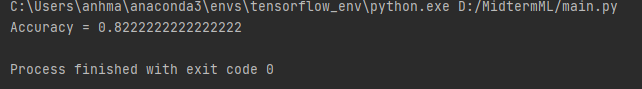
* SVM(Support vector machine): Là thuật toán học giám sát được sử dụng cho phân lớp dữ liệu.
* Với bài toán phân chia dữ liệu ta cần dung thuật toán để ánh xạ bộ data vào không gian nhiều chiều, từ đó tìm ra siêu mặt phẳng để phân chia.
* Với bài toán phân lớp dữ liệu ta sử dụng thư viện SVC của SVM sklearm.
  + Đầu tiên ta mã hóa bộ văn bản để train thành một ma trận vector
  + Sau đó trích xuất đặc trưng bằng TF-IDF
  + Set gamma = auto(scale)
* Bắt đầu train:



1.3.4 Cách đánh giá

-Acurracy:





**1.4 Thực nghiệm**

1.4.1 Công nghệ sử dụng

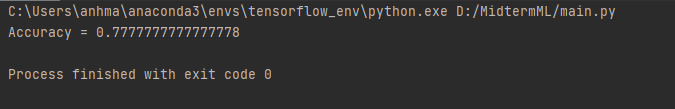
* Các thư viện sử dụng:
  1. Numpy
  2. Request
  3. Bs4 – BeautifulSoup
  4. Pandas
  5. Sklearn
  6. Underthesea

1.4.2 Các tham số của mô hình

* Vector số của văn bản để train
* TF-IDF: Đặc trưng của mô hình
* SVC

1.4.3 Kết quả đạt được

* Sau các lần train kết quả đạt được: Acurracy: 77,7%



1.4.4 Thảo luận

Phân tích kết quả đạt được và thảo luận

**TỰ ĐÁNH GIÁ**

|  |  |  |
| --- | --- | --- |
| **Nội dung** | **Điểm** | **Ghi chú** |
| 1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu **(2 điểm)** | **2** | **Lấy được toàn bộ dữ liệu theo yêu cầu** |
| 1.2 Sơ đồ tổng quát **(2 điểm)** | **1.5** |  |
| 1.3 Đặc trưng của mô hình **(3 điểm)** |  |  |
| 1.3.1. Tiền xử lý **(0.5 điểm)** | **0.5** |  |
| 1.3.2 Feature **(1 điểm)** | **1** |  |
| 1.3.3 Giải thuật **(1 điểm)** | **0.5** |  |
| 1.3.4 Cách đánh giá **(0.5 điểm)** | **0.5** |  |
| 1.4 Thực nghiệm **(3 điểm)** |  |  |
| 1.4.1 Công nghệ sử dụng **(0.5 điểm)** | **0.5** |  |
| 1.4.2 Các tham số của mô hình **(0.5 điểm)** | **0.5** |  |
| 1.4.3 Kết quả đạt được **(1.25 điểm)** | **0.75** |  |
| 1.4.4 Thảo luận **(0.75 điểm)** |  |  |
| **Tổng điểm** | **7.75** |  |

**TÀI LIỆU THAM KHẢO**

1. <https://nguyenvanhieu.vn/tf-idf-la-gi/>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
3. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>
4. <https://vi.wikipedia.org/wiki/Tf%E2%80%93idf>