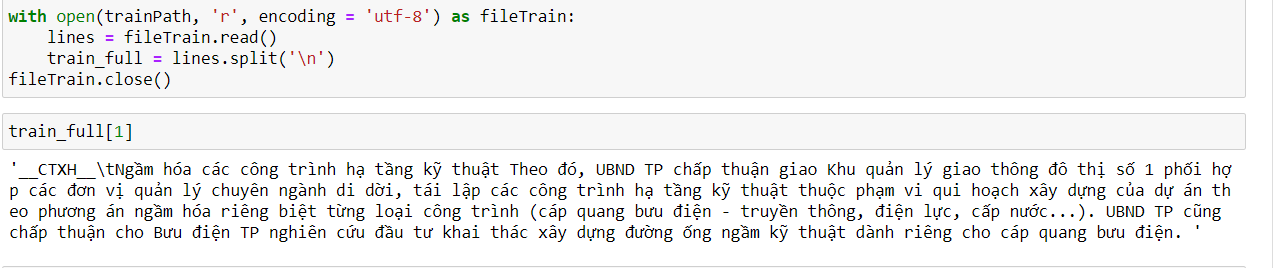
BÁO CÁO PHÂN LỚP VỚI TẬP DỮ LIỆU 2 : NLP TASK

Họ tên : Phạm Văn Sang

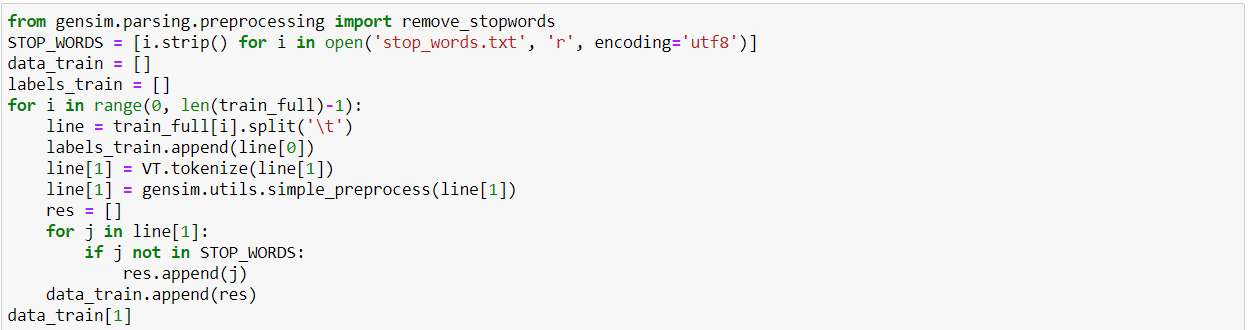
Mã sinh viên : 19020416. Nhóm : 3.

1. Dữ liệu và tiền xử lý:

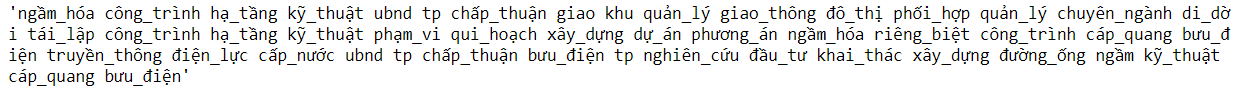
* Train data: tập dữ liệu tiếng Việt, 5000 văn bản, với 10 nhãn.
* Test data: tập dữ liệu tiếng Việt, 5000 văn bản, k có nhãn. => cần phân lớp
* Bài toán : Document/ Text classification là 1 phần điển hình trong Supervised Learning, có nhiều ứng dụng: phân loại email spam, sentiment analysis, tạo chatbot….
* Tiền xử lý: Sử dụng Bag of Words hoặc TF\_IDF. Các bước :
* Làm sạch : loại bỏ noise(Ví dụ: các thẻ HTML, JS….) => có thể sử dụng RE hoặc Gensim.
* Đọc dữ liệu từ file trainning.txt và testing.txt => tách các văn bản theo từng dòng. Với dữ liệu train:

Tương tự với dữ liệu test

* Chuẩn hóa từ : chuẩn hóa dạng viết hoa, không viết hoa, khác font…
* Tách từ trong câu : Với tiếng Việt, dấu cách kh được sử dụng như 1 kí hiệu tách từ như là các ngôn ngữ khác. Ví dụ: từ “chấp thuận” k thể dung dấu cách để tách vì nghĩa của “chấp” và “thuận” k hề liên quan đến nghĩa của từ ban đầu. => Sử dụng thư viện Pyvi của Python để tách.
* Loại bỏ Stop\_word : Là những từ xuất hiện nhiều trong các văn bản => kh có giá trị trong phân loại văn bản. Ví dụ: là, để, này,….



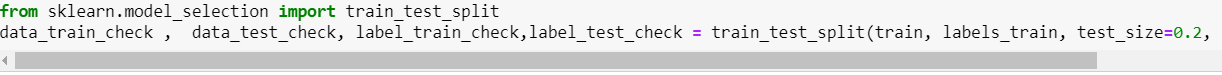
Kết quả thu được sau khi tách từ, loại bỏ các ký tự đặc biệt, stop\_word và chuyển sang string:



* Vector hóa văn bản sau các bước trên để đưa vào mô hình. Sử dụng TF-IDF :

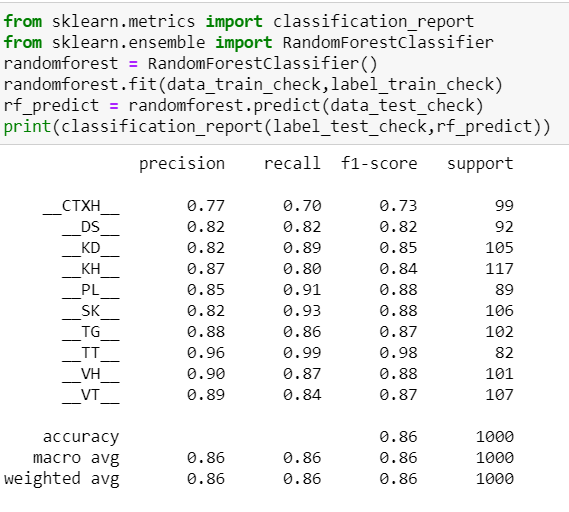
1. Xây dựng mô hình :

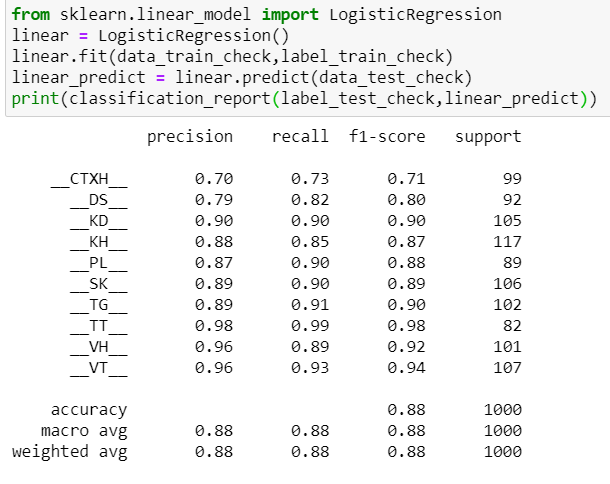
Chia tập dữ liệu train thành 2 phần để huấn luyện và đánh giá.(80/20).



Sử dụng các thuật toán có sẵn của thư viện sklearn để xây dựng mô hình và đánh giá :

* Với Random Forest:



* Với Linear Regression: 

1. Sử dụng mô hình dự đoán và xuất ra file:

