**Phân loại văn bản.**

1. **Chuẩn bị dữ liệu:**

* Dữ liệu đầu vào:

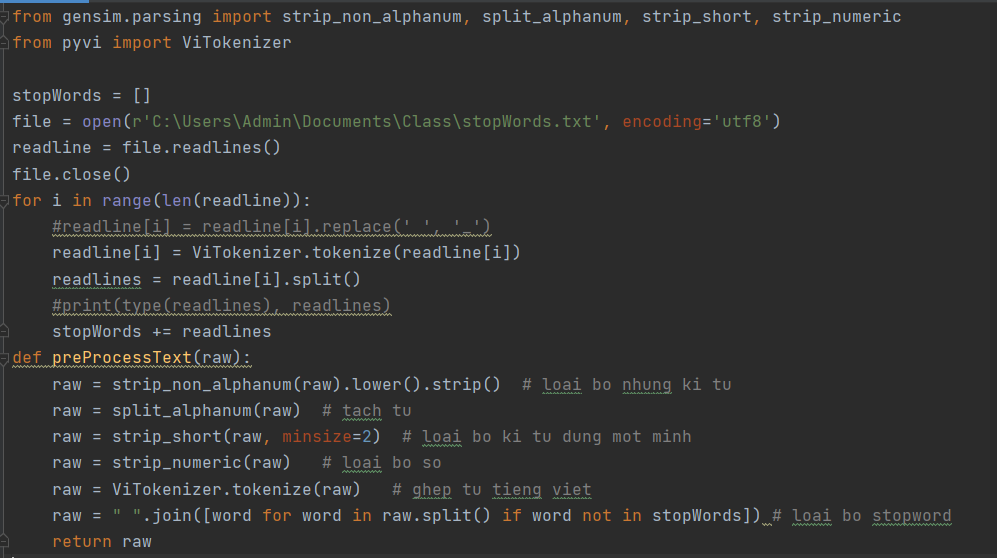
Datatraning: gồm các chủ đề sau:

* Kinh doanh
* Pháp luật
* Sức khỏe
* Thể thao
* Vi tính

Mỗi thể loại gồm khoảng 700 file.

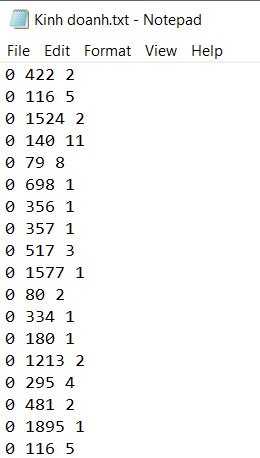
* Xử lí dữ liệu
* Loại bỏ kí tự
* Tách từ tiếng việt
* Loại bỏ kí tự đứng 1 mình
* Loại bỏ số
* Ghép các từ tiếng việt( ghép thành 1 cụm có nghĩa)
* Loại stopword

Hàm preprocess:



Hàm preprocessData: xử lý dử liệu thô từ TrainFull rồi lưu vào DataTrain

* Tạo tập từ điển là 1 danh sách gồm 2500 từ xuất hiện nhiều nhất trong dữ liệu đào tạo
* Từ dữ liệu văn bản đầu vào cần test ta tạo ra vector tương ứng như sau.



Ý nghĩa của từng dòng là:

Ví dụ: 0 422 2 có nghĩa là từ thứ 422 trong từ điển xuất hiện 2 lần trong văn bản thứ 0

1. **Tính toán.**

Hàm calculate: tính toán các xác suất. Dùng dữ liệu của DataTrainFeature nạp vào ma trận rồi thực hiện tính toán.

Cách tính theo công thức Navie Bayes có một hạn chế là nếu có 1 từ mới xuất hiện trong test nhưng không tồn tại trong tập từ điển thì xác xuất từ ấy xuất hiện trong văn bản bằng 0 => P(chủ đề| X) = 0 dù giá trị các xác suất khác có giá trị như thế nào.

* Để giải quyết vấn đề này ta sử dụng kĩ thuật laplace smoothing

*P*(*xi*​| chủ đề C ​)=(*tần xuất xuất hiện của xi trong các văn bản chủ để C* + 1)

*/* ( số các từ có trong văn bản chủ đề C + số từ có trong từ điển )

Hàm filter: tính xác suất của các chủ đề trong trường văn bản đầu vào. Đưa ra kết quả.