Bộ phân loại naive Bayes cho phân loại văn bản 2 lớp và nhiều lớp

*Đầu vào: tập văn bản tiếng Anh lấy tầm 5000 văn bản từ trang báo điện tử tiếng Anh theo chủ đề*

* *Khoa học*
* *Chính trị*
* *Kinh tế*
* *Giải trí*
* *Thể thao*

*Mỗi chủ đề tầm 1000 bài*

*URL -> crawl về, sau đó bóc tách lấy nội dung và lưu vào file*

*Phần lấy file làm chung và chia sẻ*

Hoặc <https://www.kaggle.com/yufengdev/bbc-text-categorization>

Dùng bộ phân loại naive Bayes có sẵn trong thư viện nếu có thể

Biểu diễn văn bản theo vector

* Vector nhị phân, không quan tâm đến tần số xuất hiện của từ
* TF/IDF với tần số xuất hiện của từ được xét

1. Văn bản → tách ra các từ
2. Từ tập từ → xây dựng từ điển (loại bỏ stopwords)
3. Biểu diễn lại văn bản dưới dạng vector (số chiều = kích thước từ điển)
4. Tính các xác suất tiền nghiệm và hậu nghiệm theo bayesian

Cần làm

1. Test thử phân loại với 1 lớp (VD. chọn lớp chủ đề khoa học và không thuộc lớp khoa học) với kích thước tập mẫu khác nhau để đánh giá độ chính xác. VD mẫu 100 và 200
2. Test phân loại nhiều lớp (5 lớp) với kích thước VD 100 và 200 mẫu

Đánh giá độ chính xác của mô hình trên bộ dữ liệu test

