Lọc văn bản theo chủ đề

* Xử lý dữ liệu thô
* Loại bỏ các kí tự
* Chuyển về dạng chữ thường
* Tách từ
* Loại bỏ kí tự đứng một mình
* Loại bỏ số
* Ghép từ tiếng việt
* Loại bỏ Stopword

+) ***Dữ liệu thô***

Nông - thuỷ sản VN đã có chỗ đứng ở Nhật

Ông Nguyễn Trung Dũng, Tham tán thương mại Đại sứ quán VN tại Nhật Bản, cho biết một số hàng nông - thủy sản đã đứng vững trên thị trường này. VN hiện là một trong 10 quốc gia hàng đầu về xuất khẩu hải sản và nằm trong 4 nước về xuất khẩu tôm, mực vào xứ sở hoa anh đào.

Lượng tôm xuất khẩu VN sang thị trường Nhật Bản đã chiếm tới 20% thị phần nhập khẩu của Nhật Bản với kim ngạch lên tới 36,5 tỷ yên, vượt dầu thô vốn là một trong những mặt hàng có kim ngạch xuất khẩu sang Nhật lớn nhất từ trước tới nay.

+) ***Dữ liệu sau khi xử lý***

nông thuỷ\_sản vn chỗ\_đứng nhật nguyễn trung\_dũng tham\_tán thương\_mại đại\_sứ\_quán vn nhật hàng nông thủy\_sản đứng vững thị\_trường vn hiện quốc\_gia hàng\_đầu xuất\_khẩu hải\_sản nằm xuất\_khẩu tôm mực xứ\_sở hoa đào tôm xuất\_khẩu vn thị\_trường nhật chiếm thị\_phần nhập\_khẩu nhật kim\_ngạch tỷ yên dầu\_thô vốn mặt\_hàng kim\_ngạch xuất\_khẩu nhật mặt\_hàng truyền\_thống cà\_phê gạo dầu\_thô than\_đá cao\_su tăng\_trưởng ổn\_định dũng khuyến\_cáo doanh\_nghiệp dệt may vn tích\_cực nâng chất\_lượng cải\_tiến mẫu\_mã giảm\_giá\_thành sản\_phẩm cạnh\_tranh ổn\_định thị\_trường phi hạn\_ngạch tiềm\_năng dũng nhu\_cầu rau nhật doanh\_nghiệp vn mở\_rộng thị\_trường hiện xuất\_khẩu rau nhật triệu usd tổng\_số tỷ usd kim\_ngạch nhập\_khẩu mặt\_hàng nhật\_bản

* Xây dựng từ điển là 1 list các từ có tần suất xuất hiện nhiều nhất trong dữ liệu đào tạo.
* Tạo các đặc trưng đầu vào của từng văn bản

Mỗi đặc trưng là 1 list có độ dài bằng với độ dài của từ điển. Trong đó giá trị thứ i của list là tần suất từ thứ i trong từ điển xuất hiện trong văn bản.

* Tính xác suất văn bản thuộc vào các chủ đề theo công thức bayes.

*Lưu ý:*

*Cách tính theo công thức Naive Bayes có một hạn chế là nếu có một từ mới chưa bao giờ xuất hiện trong từ điển thì biểu thức xác suất từ ấy xuất hiện trong văn bản sẽ bằng không, dẫn đến xác suất* **p(chủ đề| X) = 0** *bất kể các giá trị còn lại có lớn thế nào. Để giải quyết việc này, một kỹ thuật được gọi là Laplace smoothing được áp dụng:*

**p(xi | chủ đề c) = (tần suất của xi trong các văn bản chủ đề c + 1) / (số các từ có trong các văn bản chủ đề c + số từ trong từ điển)**

*Các xác suất thường rất nhỏ nên thay vì tính tích các xác suất thì thay bằng tính tổng các log của các xác suất thành phần (vì hàm log là hàm đồng biến trên tập số dương)*

* Chọn ra chủ đề có xác suất lớn nhất.