Bài toán phân loại tin tức trên báo điện tử

Mục tiêu của một hệ thống phân loại văn bản là nó có thể tự động phân loại một văn bản cho trước

Bài toán *phân loại văn bản* là một bài toán học giám sát (supervised learning) trong học máy (machine learning), bởi vì nội dung của văn bản đã được gán nhãn, và được sử dụng để thực hiện phân loại.

Bài toán:

- Đầu vào: Tên của một bài báo điện tử

Nguồn: VnExpress,Báo thanh niên,….

-Đầu ra: Thể loại bài báo:

Kinh tế, Thể thao, Giáo dục, Sức khỏe, Du lịch, Pháp luật,…

Vd: Tên Bài Báo: “Trung Quốc mở cửa tác động thế nào đến kinh tế toàn cầu

Thể loại:Kinh doanh

*Mô tả bộ dữ liệu:

Cách thức xây dựng: Có sẵn sử dụng crawl từ VnExpress

https://vnexpress.net/

Dữ liệu cần chuẩn bị là tên của bài báo kèm theo chủ để của bài báo đó:

Trong đó label được gán sẵn: Khoa học, tin tức, sức khỏe,….

Số lượng: 10 label

- Chính trị xã hội
- Đời sống
- Khoa học
- Kinh doanh
- Pháp luật
- Sức khoẻ
- Thế giới
- Thể thao
- Văn hoá
- Vi tính

*Tiền xử lý dữ liệu:

Gồm 2 quá trình: Chuẩn hóa dữ liệu, loại bỏ các thành phần không có ý nghĩa

- Chuẩn hóa kiểu gõ dấu tiếng việt
- Xóa các ký tự đặc biêt: “ , ” , ‘ , ; ,
- Đưa về dạng chữ thường
- Tách từ tiếng việt
- Vecto hóa

*Chuẩn hóa kiểu gõ dấu:

Kiểu gõ dấu khác nhau thì bạn nhìn mắt thường cũng sẽ thấy được sự khác nhau: òa với oà lần lượt là kiểu gõ cũ (phổ biến hơn) và kiểu gõ mới.

*Tách từ tiếng việt

Học sinh học sinh học

=> Học_sinh học sinh_học

*Đưa về viết thường

Các giai đoạn:

Giai đoạn 1:

Huấn luyện (training) là giai đoạn học tập của mô hình phân loại văn bản.

Ở bước này, mô hình sẽ học từ dữ liệu có nhãn (trong ảnh trên nhãn là Possitive, Negative, Neutral). Dữ liệu văn bản sẽ được số hóa thông qua bộ trích xuất đặc trưng (feature extractor) để mỗi mẫu dữ liệu trong tập huấn luyện trở thành 1 vector nhiều chiều (đặc trưng). Thuật toán máy học sẽ học và tối ưu các tham số để đạt được kết quả tốt trên tập dữ liệu này. Nhãn của dữ liệu được dùng để đánh giá việc mô hình học tốt không và dựa vào đó để tối ưu.

Giai đoạn 2

Dự đoán (prediction), là giai đoạn sử dụng mô hình học máy sau khi nó đã học xong. Ở giai đoạn này, dữ liệu cần dự đoán cũng vẫn thực hiện các bước trích xuất đặc trưng. Mô hình đã học sau đó nhận đầu vào là đặc trưng đó và đưa ra kết quả dự đoán.

Thuật toán sử dụng:

- Logistic Regression (thư viện sklearn)
- Support Vector Machine (thư viện sklearn)

*Cách đánh giá:

Sử dụng độ chính xác (Accuracy)