Bài toán phân loại tin tức trên báo điện tử

- Đầu vào: Tên của một bài báo điện tử

Nguồn: VnExpress,Báo thanh niên,….

-Đầu ra: Thể loại bảo báo:

Thể thao, Chính trị, Đời sống, Du lịch, vv,…

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

Thể loại:Kinh tế

*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

*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