Lúc trước 77-83%

Sau khi bỏ đi bước “loại bỏ stopword” thì độ chính xác giảm đi 1-2%

Mỗi thể loại 250 tin. 6 cate = 1500 tin

Crawl lại data train. Mỗi thể loại tầm 1000 tin. 6 cate = gần 6000 tin.

Từ 6 website đang tin cậy nhát VN hiện nay

Vs bộ train mới thì độ chính xác tăng lên đáng kể vs tỷ lệ 85-92%

Lúc test thì data có dạng như nào ?

1 mảng gồm nhiều phần tử là chuỗi đại diện cho cả 6 loại tin

Hay 6 mảng và mỗi mảng đại diện cho 1 thể loại

…

Giữ nguyên cách hiện tại đc không hay bắt buộc phải thay đổi để gán idtheloai cho mỗi tin tức trước khi phân loại, rồi đem so sánh ???

Quy trình làm:

Crawl data training 🡪 array\_training

Crawl data testing 🡪 array\_testing

Classifier 🡪 filenew 🡪 handling\_data 🡪chartJS

Có nên bỏ con số đi ko

9000 chiếc xe

Thế hệ 9x