Môn học: Khoa học dữ liệu

Đồ án cuối kỳ:
Phân loại văn bản tin tức

Sinh viên 1: Trần Quang Minh
Sinh viên 2: Đoàn Quang Tuấn

Bổ sung so với buổi vấn đáp

- Slide: Bổ sung slide số 6, kiểm tra dữ liệu thu thập là hợp lệ.
- Notebook: Bổ sung các chú thích, sửa lỗi chính tả.

Nội dung

- Phát biểu bài toán
- Thu thập dữ liệu
- Thống kê quan sát trên dữ liệu
- Tiền xử lý dữ liệu
- Chọn các đặc trưng
- Thiết kế mô hình
- Huấn luyện, đánh giá và kiểm thử
- Tổng kết

Phát biểu bài toán

Câu hỏi: Dựa vào nội dung của 1 văn bản tin tức, dự đoán xem đây là văn bản thuộc thể loại nào?

Ứng dụng:

- Tự động hóa quá trình phân loại cho các diễn đàn tin tức
- Phần nào hỗ trợ việc lọc các nội dung

Thu thập dữ liệu

Dữ liệu được lấy từ trang web: **vietnamnews.com**

Thu thập dữ liệu

Kiểm tra dữ liệu thu thập hợp lệ,
robots.txt:

```
User-agent: *
Allow:/
Disallow: /ajax/
Disallow: /*.ashx/
Disallow: /Scripts/
Disallow: /Resource/
Disallow: /serviceadv/
Disallow: /bizhub/
Disallow: /ovietnam/
Sitemap: http://vietnamnews.vn/sitemap.xml
```

Thu thập dữ liệu

Sử dụng 5 url có sẵn của 5 thể loại:

- Politics-laws
- Society
- Economy
- Sport
- Environment

Thu thập dữ liệu

Các đường dẫn này sẽ đi đến danh sách các bài viết thuộc chủ đề.

Khó khăn: các bài viết được phân trang => cần tương tác với browser

Giải quyết: sử dụng selenium, tương tác với nút Next, và lấy url của từng bài viết.

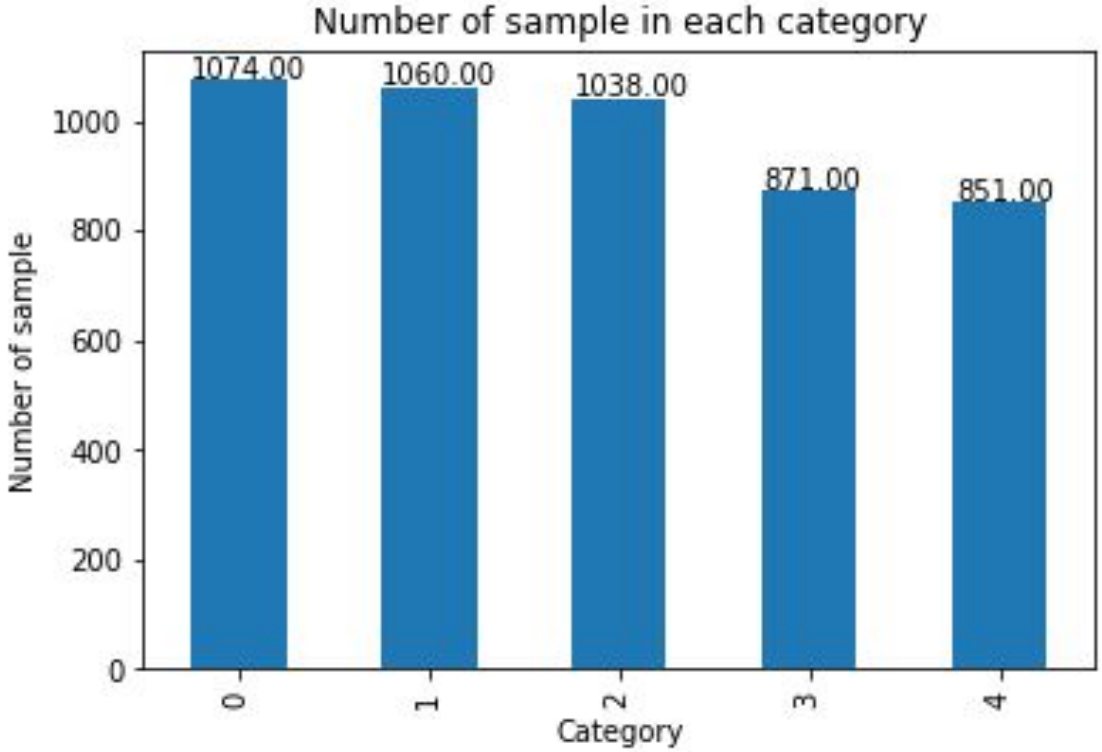
Thu thập dữ liệu

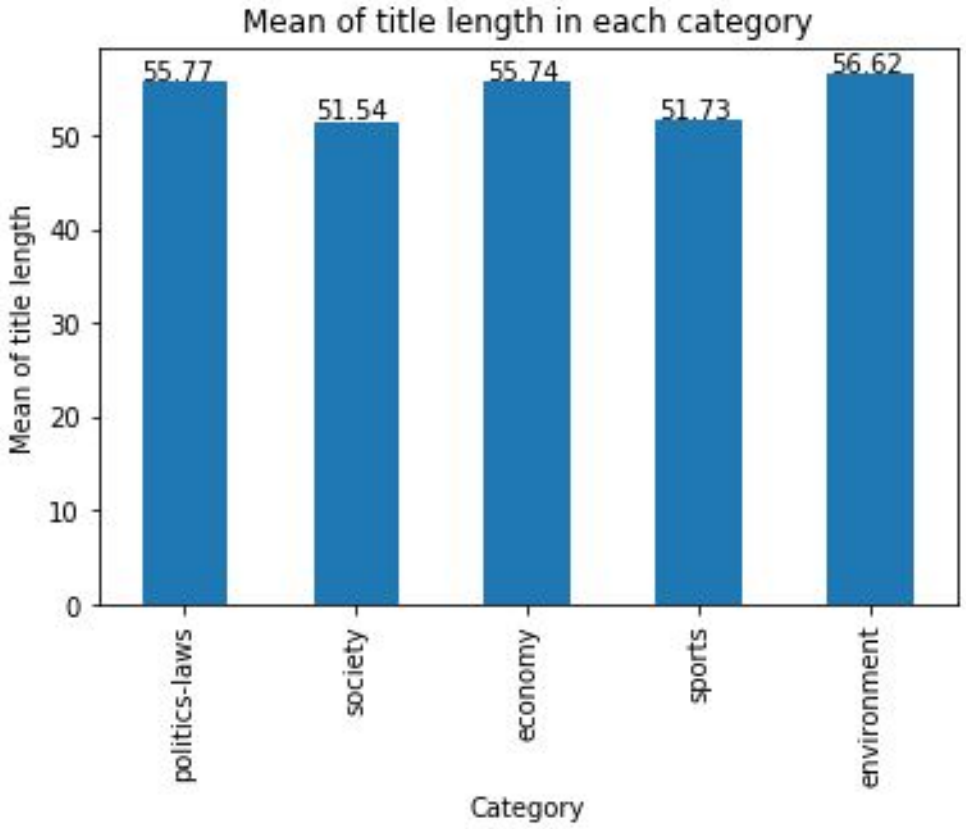
- Với mỗi thể loại, selenium tạo một driver để lấy url trên từng trang.

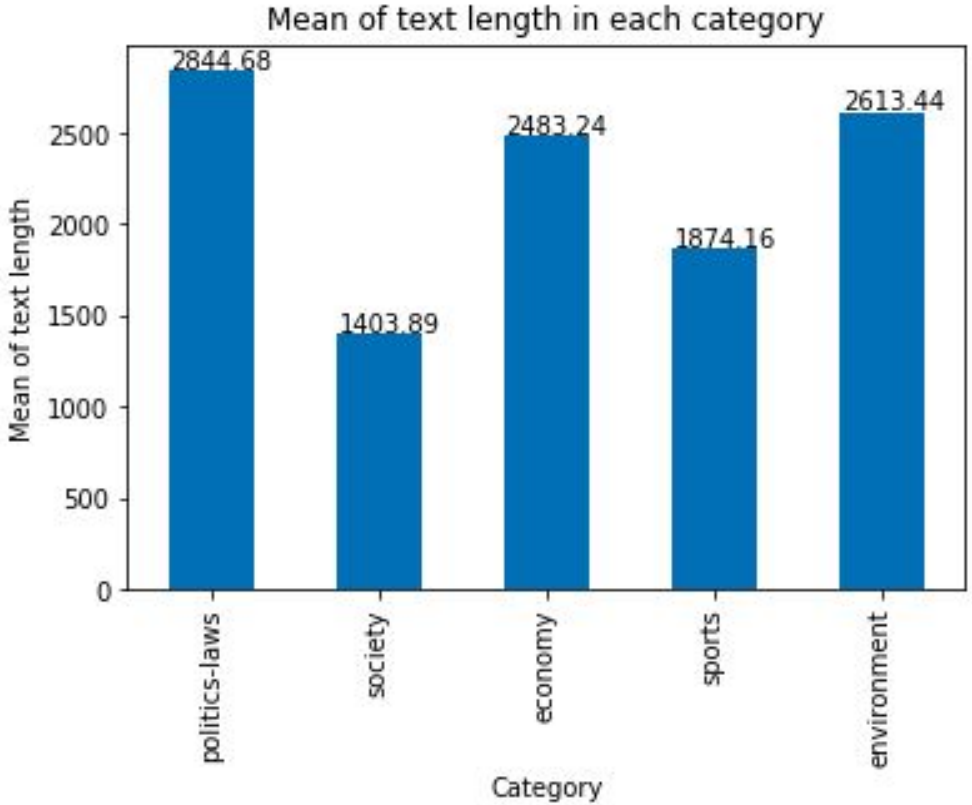
- Với mỗi url, thực hiện thu thập các thông tin về tiêu đề và nội dung văn bản, dựa vào các tiêu chí về thời gian, độ dài nội dung, tiêu đề.

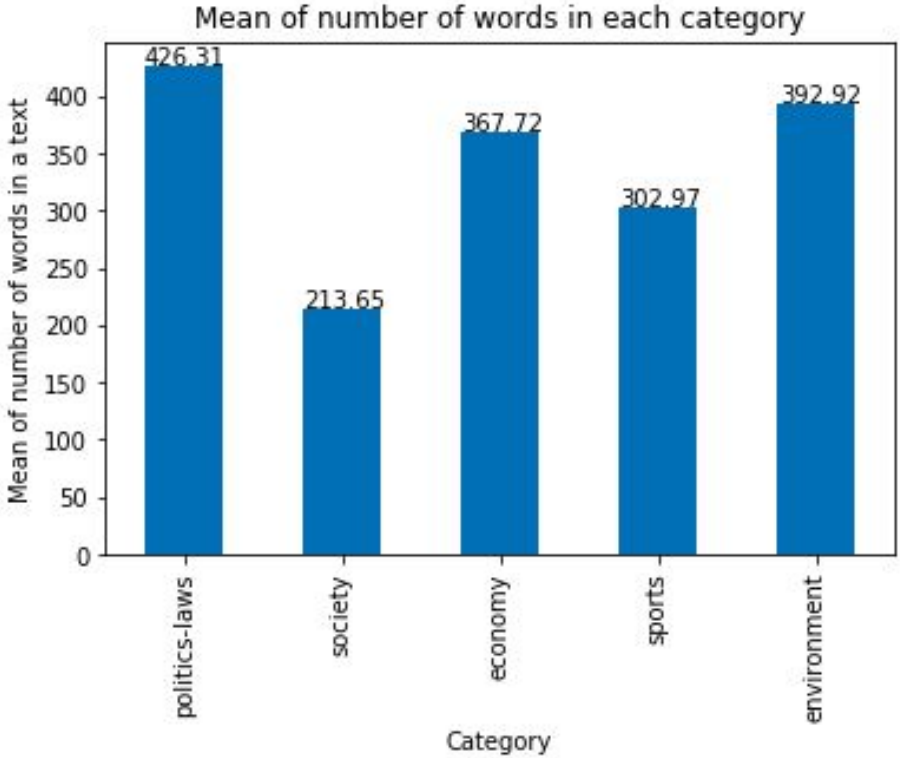
Thống kê quan sát trên dữ liệu

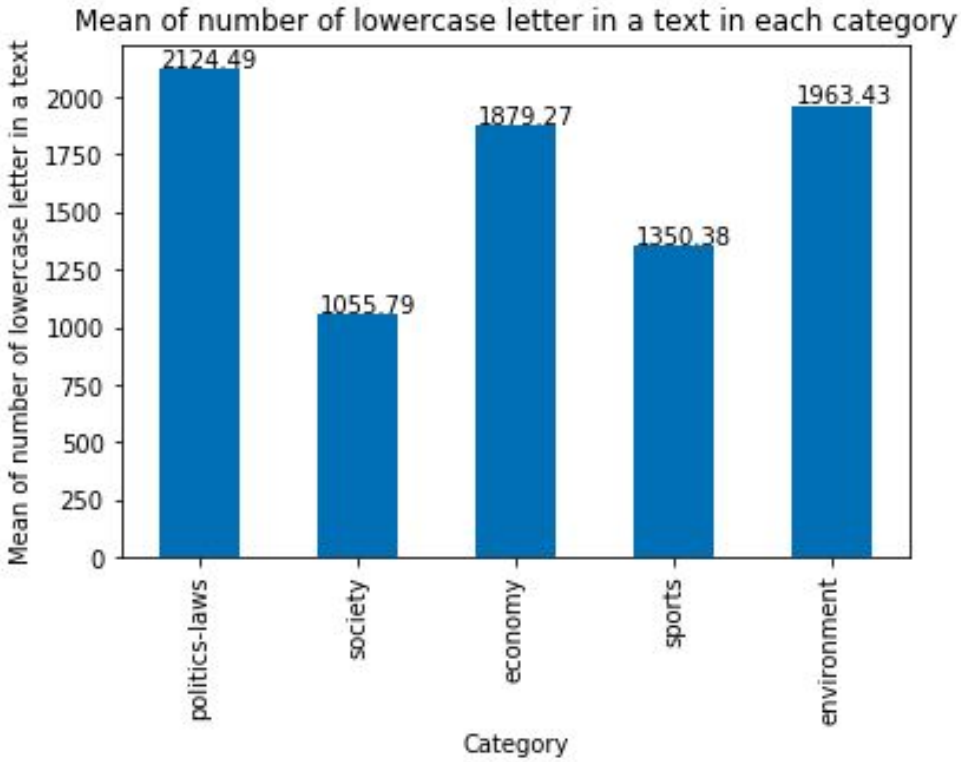
- Số lượng mẫu dữ liệu mỗi thể loại
- Độ dài chuỗi trung bình của tiêu đề (title) ở mỗi thể loại
- Độ dài chuỗi trung bình của văn bản (text) ở mỗi thể loại
- Số lượng từ trung bình trong 1 văn bản của từng thể loại
- Số lượng chữ cái viết thường trung bình trong 1 văn bản của từng thể loại
- Số lượng chữ cái viết hoa trung bình trong 1 văn bản của từng thể loại
- Số lượng dấu cảm thán (!, ?) trung bình trong 1 văn bản của từng thể loại

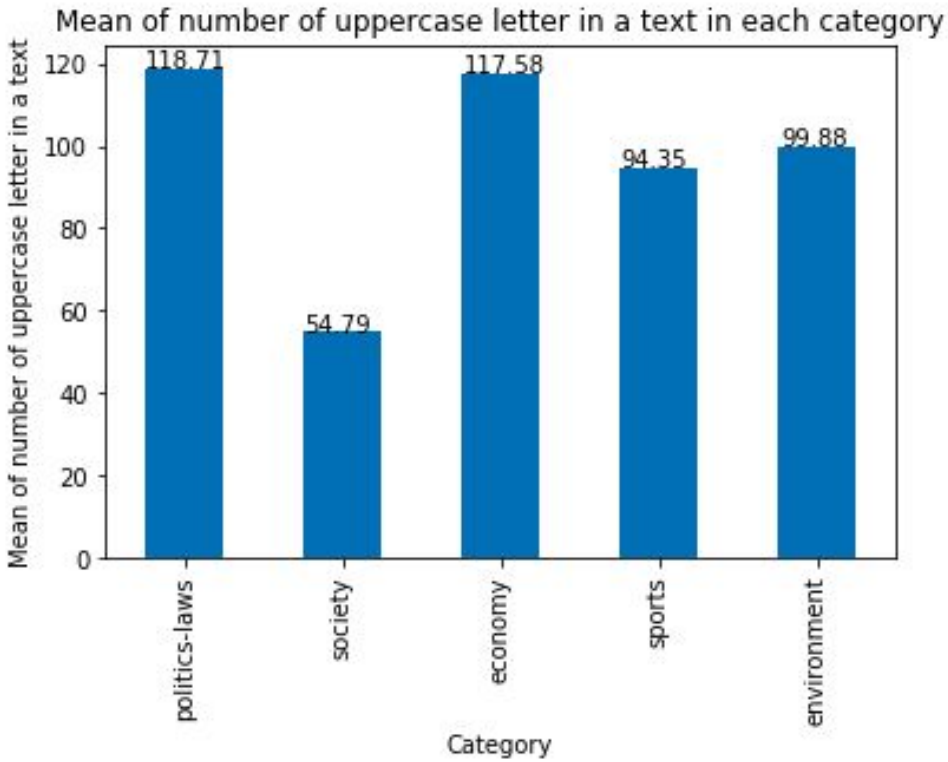

Number of sample in each category
1074.00
1060.00
1038.00
871.00
851.00
1000
800
600
400
200
0
Number of sample
0
1
2
3
4
Category

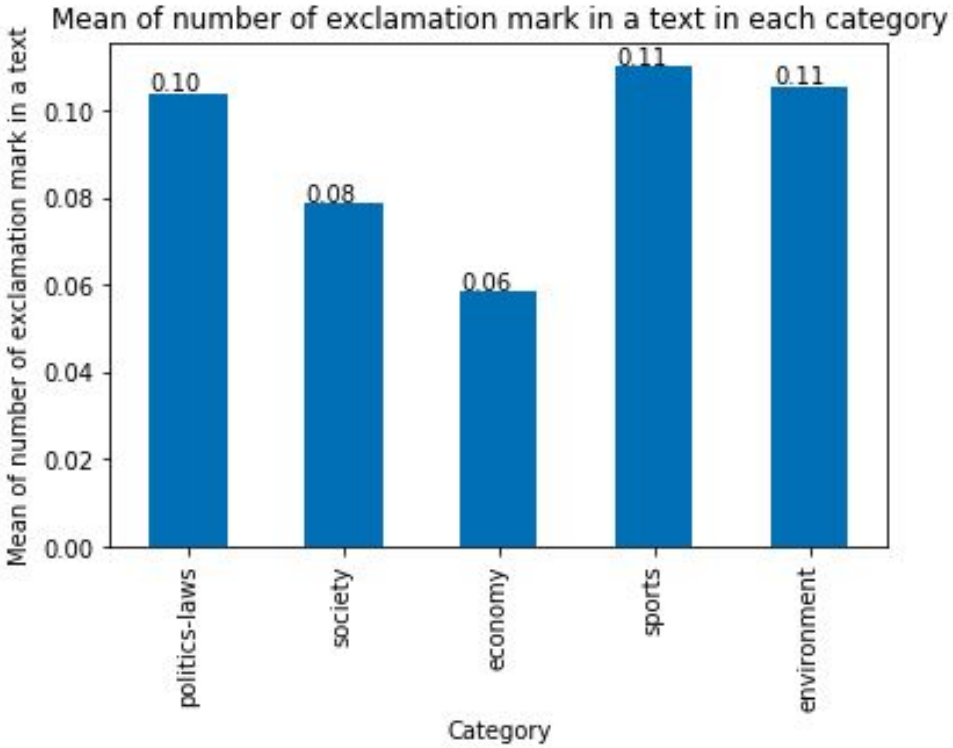

Mean of title length in each category
55.77
51.54
55.74
51.73
56.62
Mean of title length
0
10
20
30
40
50
politics-laws
society
economy
sports
environment
Category


Mean of text length in each category
Mean of text length
Category
2844.68
1403.89
2483.24
1874.16
2613.44
0
500
1000
1500
2000
2500
politics-laws
society
economy
sports
environment


Mean of number of words in each category
Mean of number of words in a text
0
50
100
150
200
250
300
350
400
426.31
213.65
367.72
302.97
392.92
politics-laws
society
economy
sports
environment
Category


Mean of number of lowercase letter in a text in each category
Mean of number of lowercase letter in a text
2124.49
1055.79
1879.27
1350.38
1963.43
2000
1750
1500
1250
1000
750
500
250
0
politics-laws
society
economy
sports
environment
Category


Mean of number of uppercase letter in a text in each category
Mean of number of uppercase letter in a text
120
100
80
60
40
20
0
118.71
54.79
117.58
94.35
99.88
politics-laws
society
economy
sports
environment
Category


Mean of number of exclamation mark in a text in each category
Mean of number of exclamation mark in a text
0.00
0.02
0.04
0.06
0.08
0.10
0.10
0.08
0.06
0.11
0.11
politics-laws
society
economy
sports
environment
Category

Tiền xử lý dữ liệu

- Loại bỏ các ký tự lỗi
- Loại bỏ các từ như 's, 've, n't, 're, 'd, 'll
- Loại bỏ các khoảng trắng thừa, ký tự xuống dòng

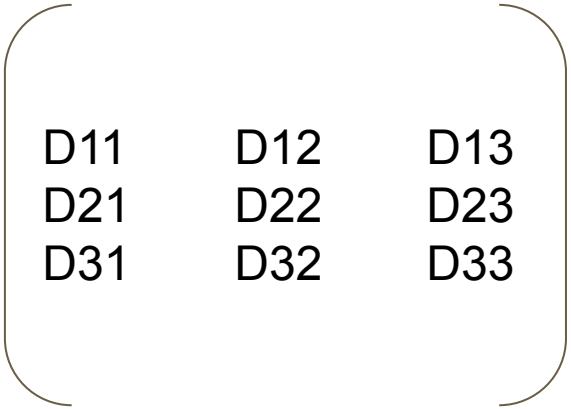
Chọn các đặc trưng

Dựa vào các quan sát về việc thống kê trên dữ liệu, các đặc trưng chung được chọn bao gồm:

- Số ký tự trong văn bản.
- Số từ trong văn bản.
- Số lượng dấu cảm thán (!, ?)
- Số chữ cái viết hoa
- Số chữ cái viết thường

Chọn các đặc trưng - TF-IDF

TF-IDF: Một đặc trưng rất thường được sử dụng khi xử lý ngôn ngữ dạng văn bản. (viết tắt của: Term Frequency - Inverse Document Frequency)



Tập từ vựng:
[A1, A2, A3, A4]

Với Dij là từ thứ j trong document thứ i

Chọn các đặc trưng - TF-IDF

Kết quả thu được sau khi chọn đặc trưng **tf-idf**:

tf-idf(A1, D1, D)	tf-idf(A2, D1, D)	tf-idf(A3, D1, D)	tf-idf(A4, D1, D)
tf-idf(A1, D2, D)	tf-idf(A2, D2, D)	tf-idf(A3, D2, D)	tf-idf(A4, D2, D)
tf-idf(A1, D3, D)	tf-idf(A2, D3, D)	tf-idf(A3, D3, D)	tf-idf(A4, D3, D)

Chọn các đặc trưng - TF-IDF

Cách tính:

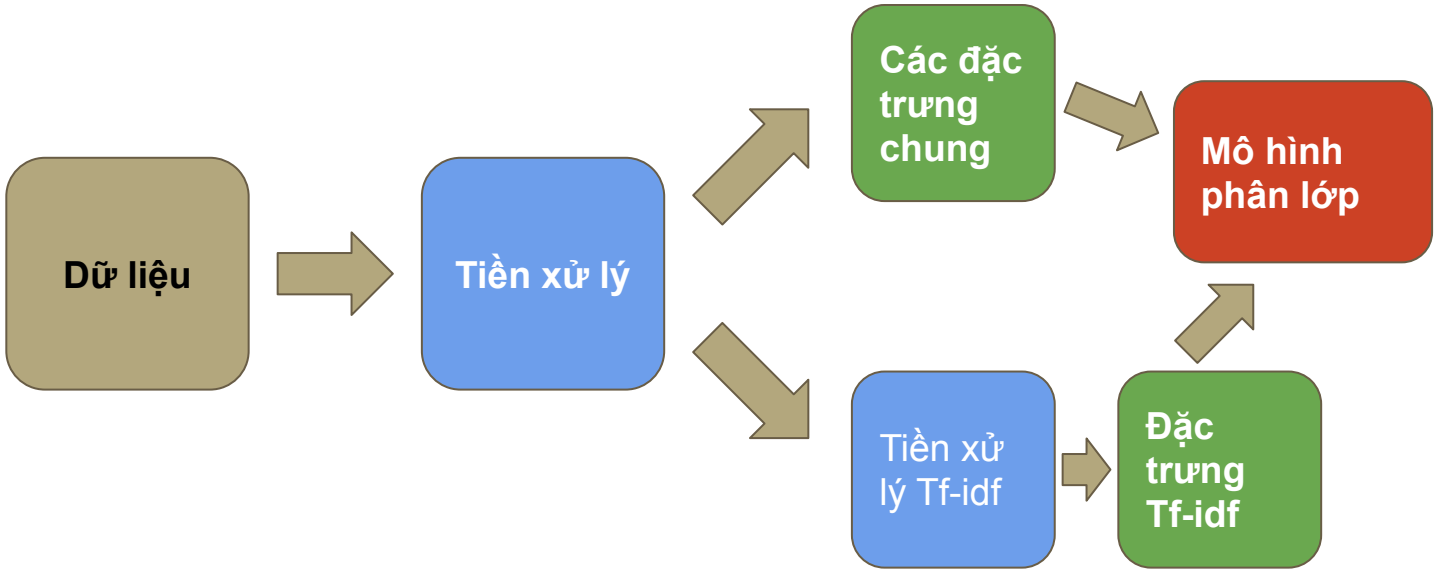
tf-idf(A, Di, D) = tf(A, Di) * idf(A, D)

Trong đó:

tf(A, Di) = Số lần xuất hiện của **từ A** trong **document Di**

idf(A, D) = ln(Tổng số document **trong D** / Số document **trong D** mà **từ A** xuất hiện)

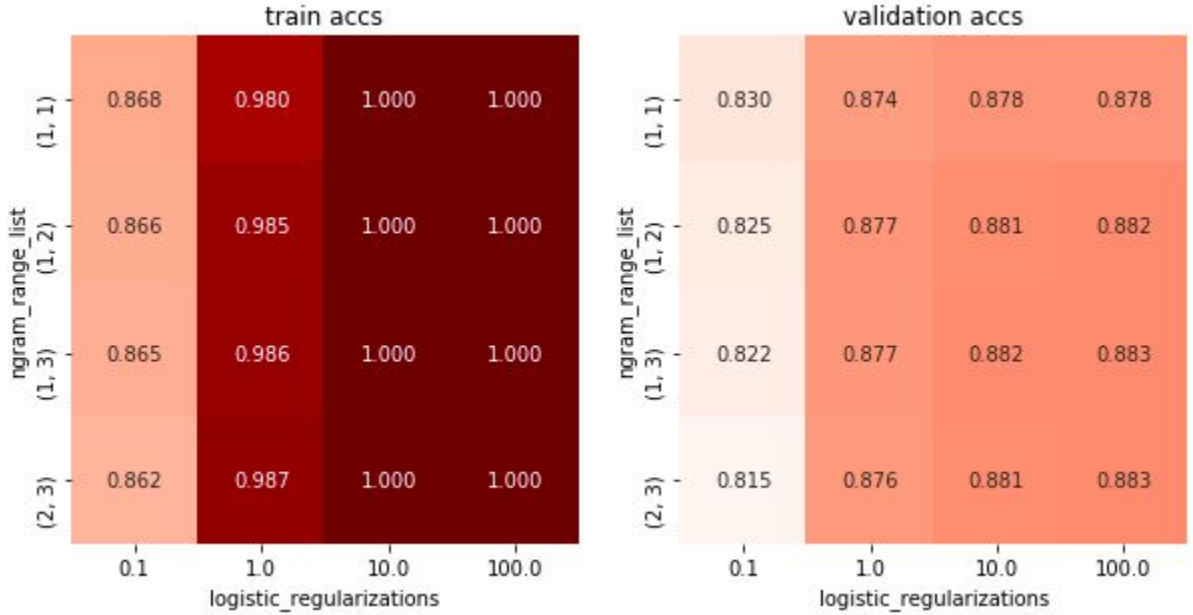
Thiết kế mô hình



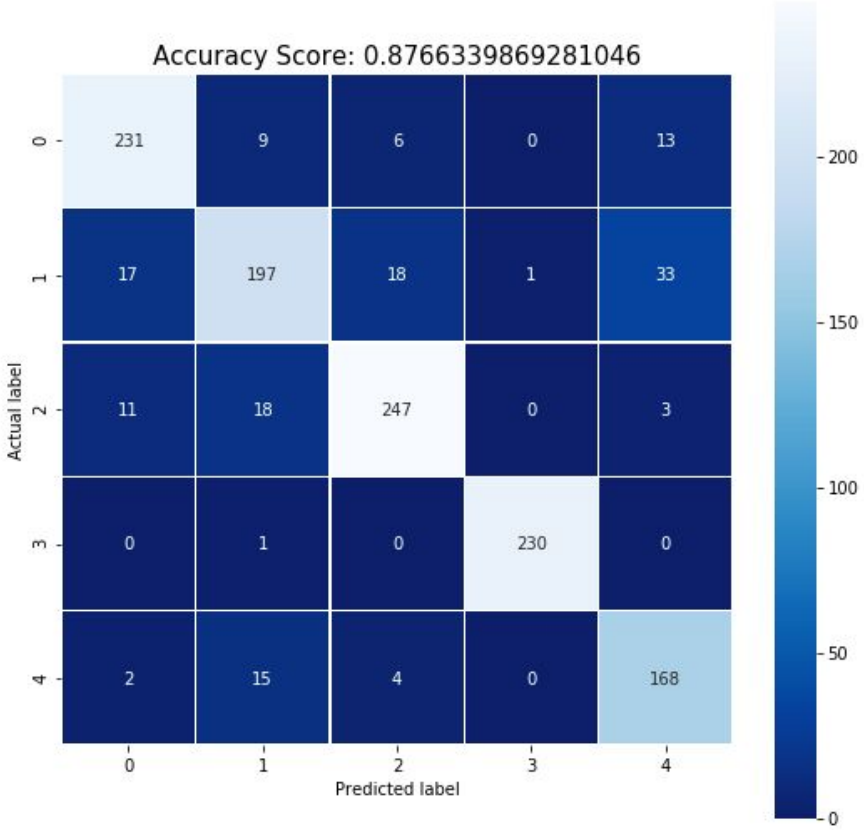
Huấn luyện, đánh giá và kiểm thử

- Mô hình: **Logistic Regression**
- Chia 5 fold để đánh giá độ chính xác
- Các thông số cần tinh chỉnh:
    + ngram_range: (1,1), (1,2), (1,3), (2,3)
    + Mức độ regularization C: 0.01, 0.1, 1, 10

Huấn luyện, đánh giá và kiểm thử



Huấn luyện, đánh giá và kiểm thử



Tổng kết - Đã làm được gì? Học được gì?

- Cách crawl dữ liệu hiệu quả
- Sử dụng một số đặc trưng cơ bản của nội dung văn bản để phân loại văn bản
- Sử dụng pipeline để huấn luyện và kiểm thử trên các mô hình
- Sử dụng, tổ chức lưu trữ với github

Hỏi - đáp