

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

# 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](http://vietnamnews.com)**



# 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.



## More work needed to fight trade fraud, smuggling: GDC

More measures are needed in the fight against trade fraud and illegal cross-border smuggling between Việt Nam and other countries, which

continues to be a problem, the General Department of Customs has said.



## Thousands of cancer patients concerns over drug shortage

Over the past three years, N T H, a 38-year-old patents from Mekong Delta of An Giang, regularly visits the National Institute of Hematology and

Blood Transfusion for chronic myeloid leukemia treatment.



## Doctors bring smiles to patients with horrible burns

Nguyễn Thị Sáu, 54 and Lê Thị Lan Vy, 24, are two of many patients who have received support from doctors of the Centre for Plastic and Reconstructive Surgery to regain their

confidence.

# Thu thập dữ liệu

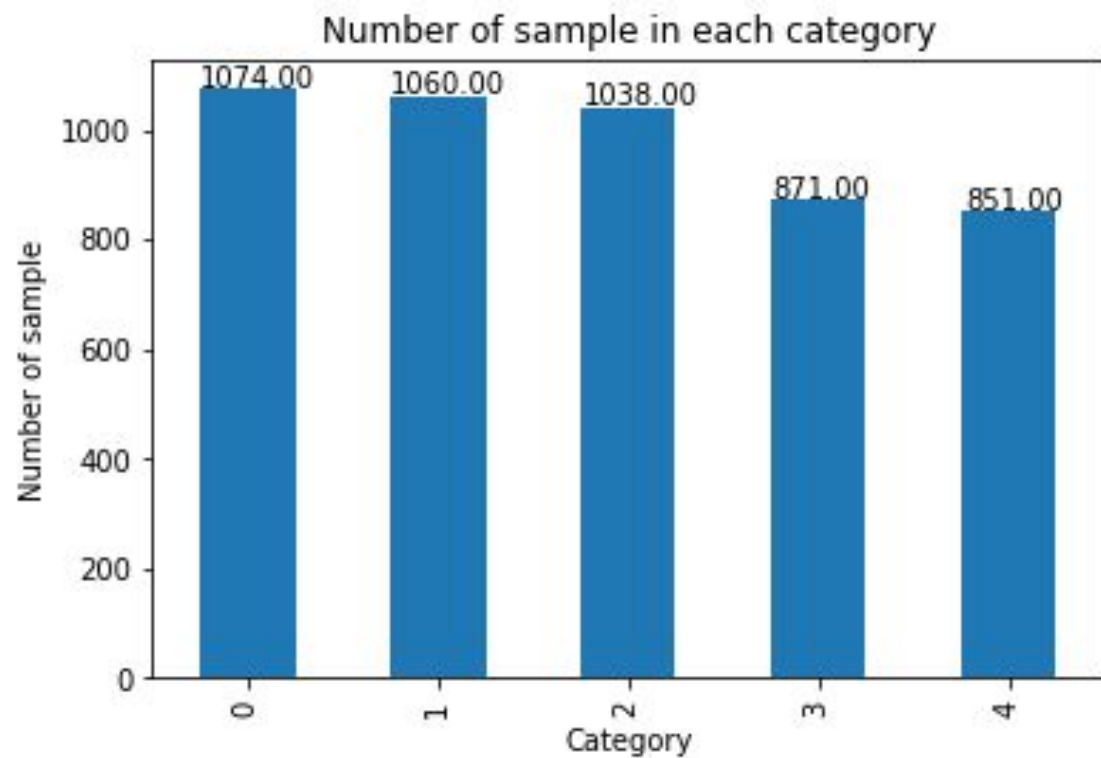
- Với mỗi thể loại, selenium tạo một driver để lấy url trên từng trang.
- Với mỗi url, thực hiện thu thập các thông tin về tiêu đề và nội dung văn bản, dựa vào các tiêu chí về thời gian, độ dài nội dung, tiêu đề.

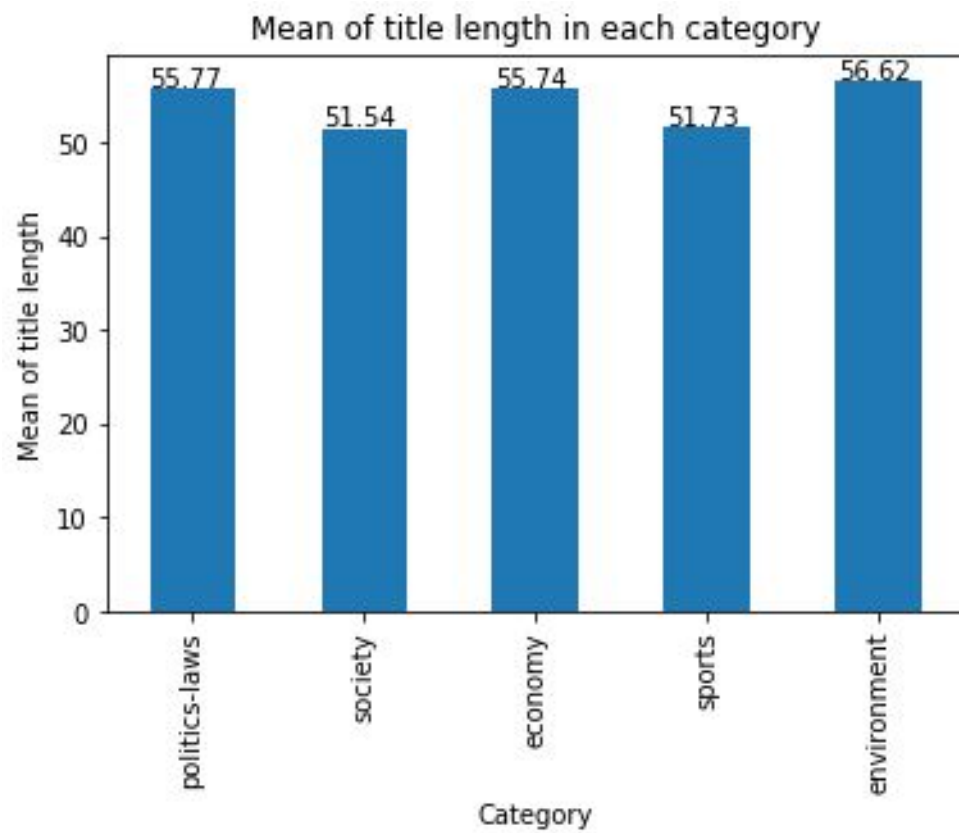
```
▶ <h3 class="vnnews-tt-post">...</h3>
▶ <div class="vnnews-time-post">...</div>
  <script type="text/javascript">categoryid = 'subcate101';</script>
▶ <div class="vnnews-text-post">...</div>
▶ <div class="vnnews-ft-post">...</div>
▶ <div style="position: relative; padding-top: 10px;">...</div>
▶ <div style="margin-top: 5px; margin-bottom: 5px;">...</div>
▶ <div class="vnnews-grp-item-cmt" style="display: none;">...</div>
▶ <div class="vnnews-box-cmt" style="margin: 16px 0 !important;">...</div>
▶ <div class="vnnews-lates-post">...</div>
...
```

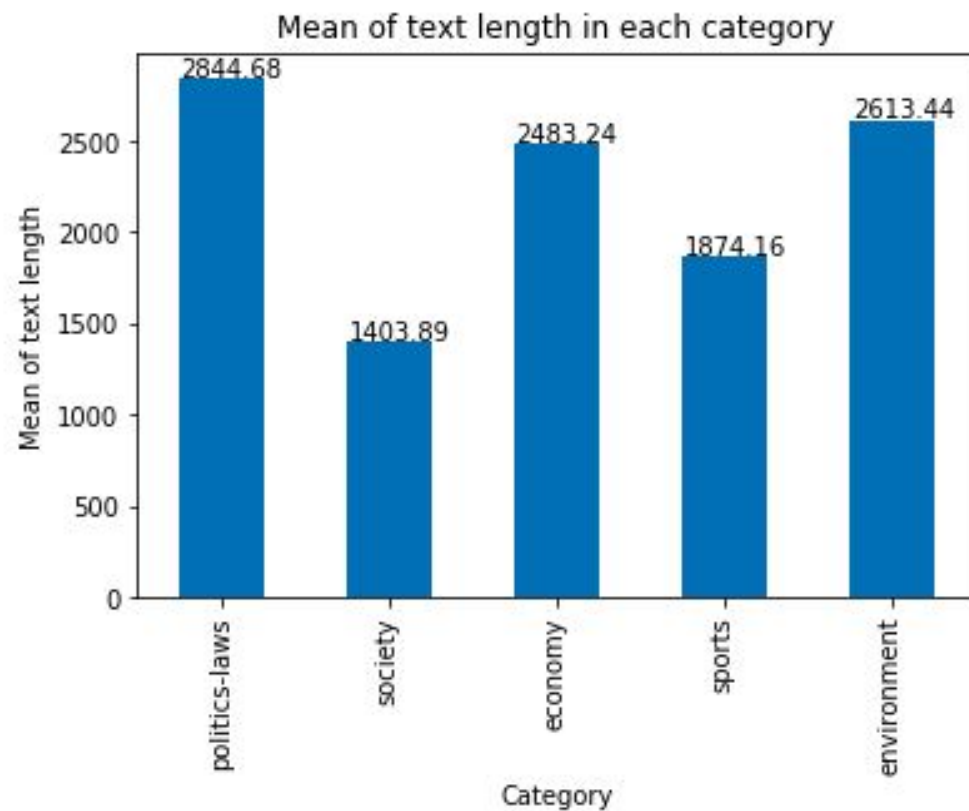
# Thống kê quan sát trên dữ liệu

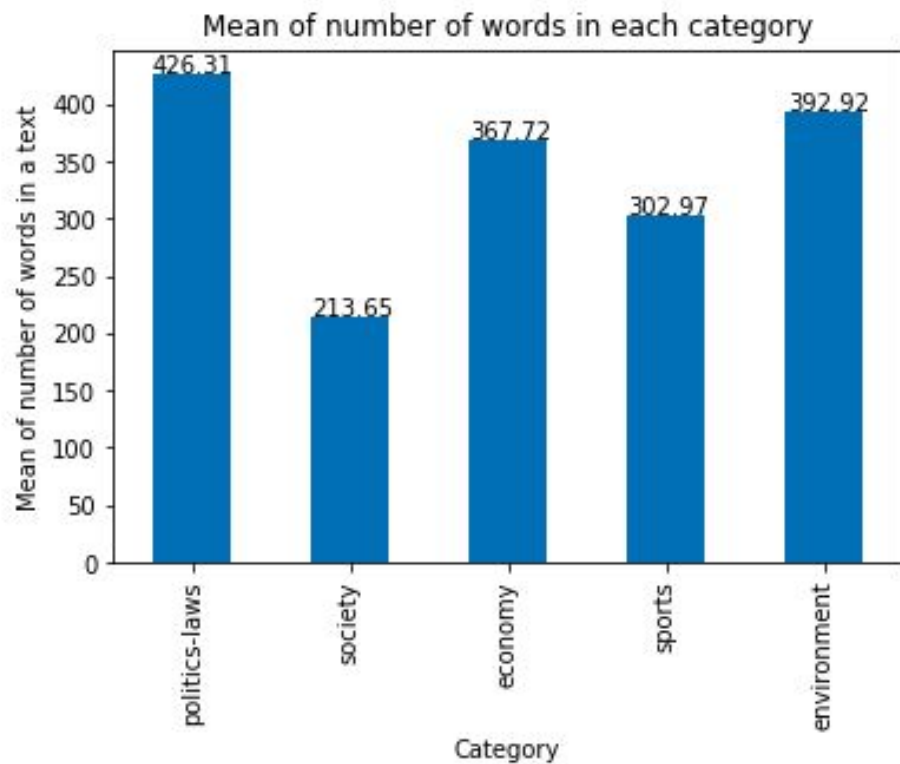
- Số lượng mẫu dữ liệu mỗi thể loại
- Độ dài chuỗi trung bình của tiêu đề (title) ở mỗi thể loại
- Độ dài chuỗi trung bình của văn bản (text) ở mỗi thể loại
- Số lượng từ trung bình trong 1 văn bản của từng thể loại
- Số lượng chữ cái viết thường trung bình trong 1 văn bản của từng thể loại
- Số lượng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

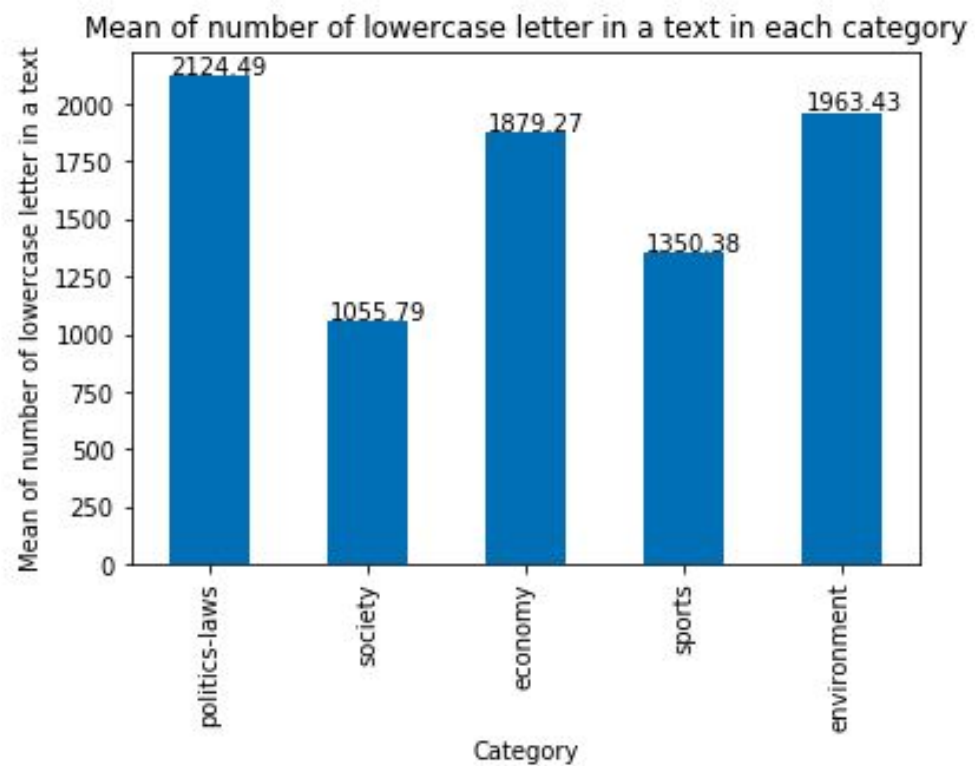


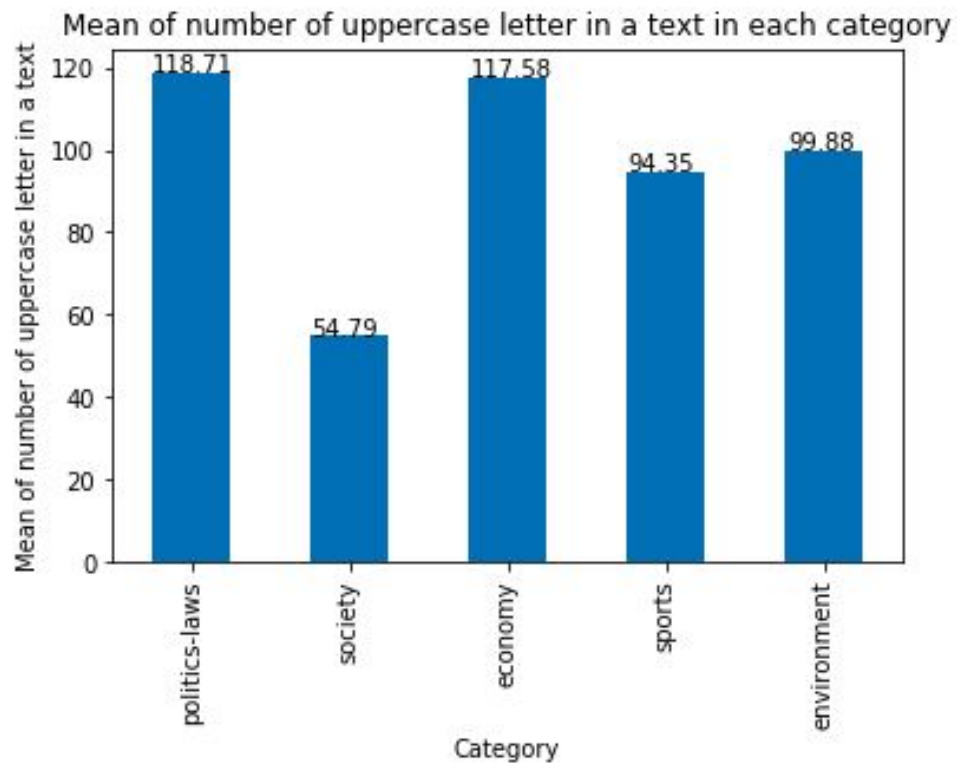


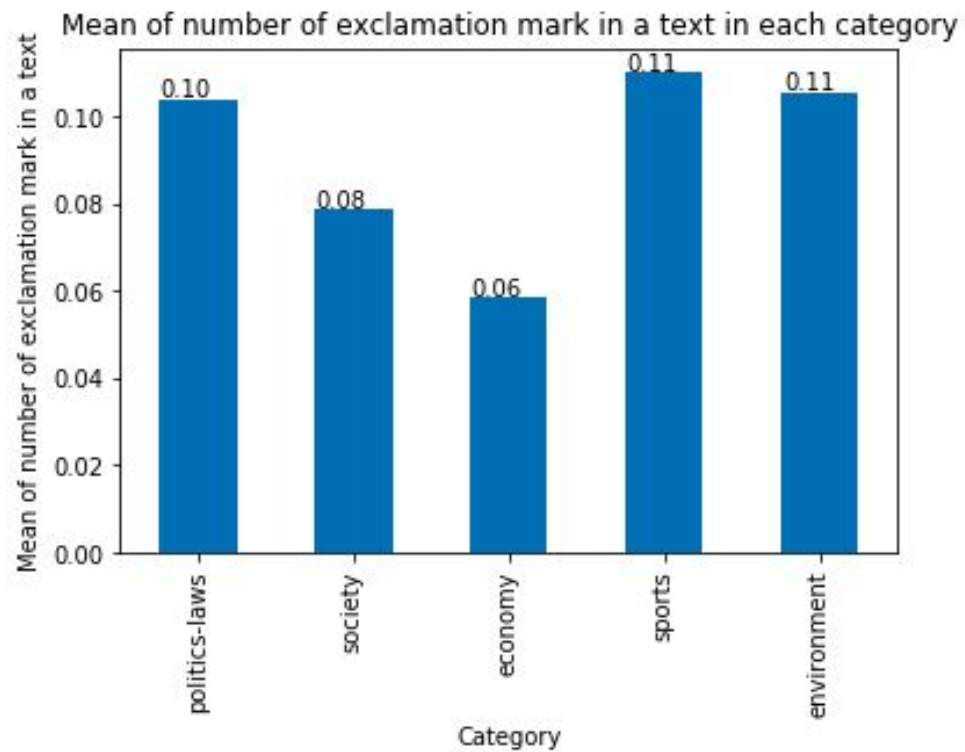












# Tiền xử lý dữ liệu

- 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



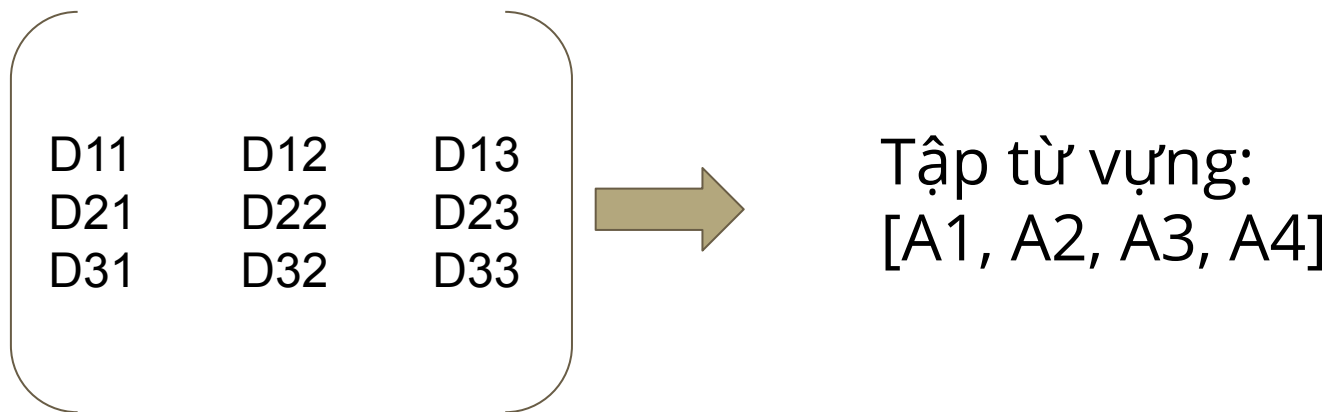
# Chọn các đặc trưng

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)



Với  $D_{ij}$  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**:

$\text{tf-idf}(A1, D1, D)$

$\text{tf-idf}(A2, D1, D)$

$\text{tf-idf}(A3, D1, D)$

$\text{tf-idf}(A4, D1, D)$

$\text{tf-idf}(A1, D2, D)$

$\text{tf-idf}(A2, D2, D)$

$\text{tf-idf}(A3, D2, D)$

$\text{tf-idf}(A4, D2, D)$

$\text{tf-idf}(A1, D3, D)$

$\text{tf-idf}(A2, D3, D)$

$\text{tf-idf}(A3, D3, D)$

$\text{tf-idf}(A4, D3, D)$

# Chọn các đặc trưng - TF-IDF

Cách tính:

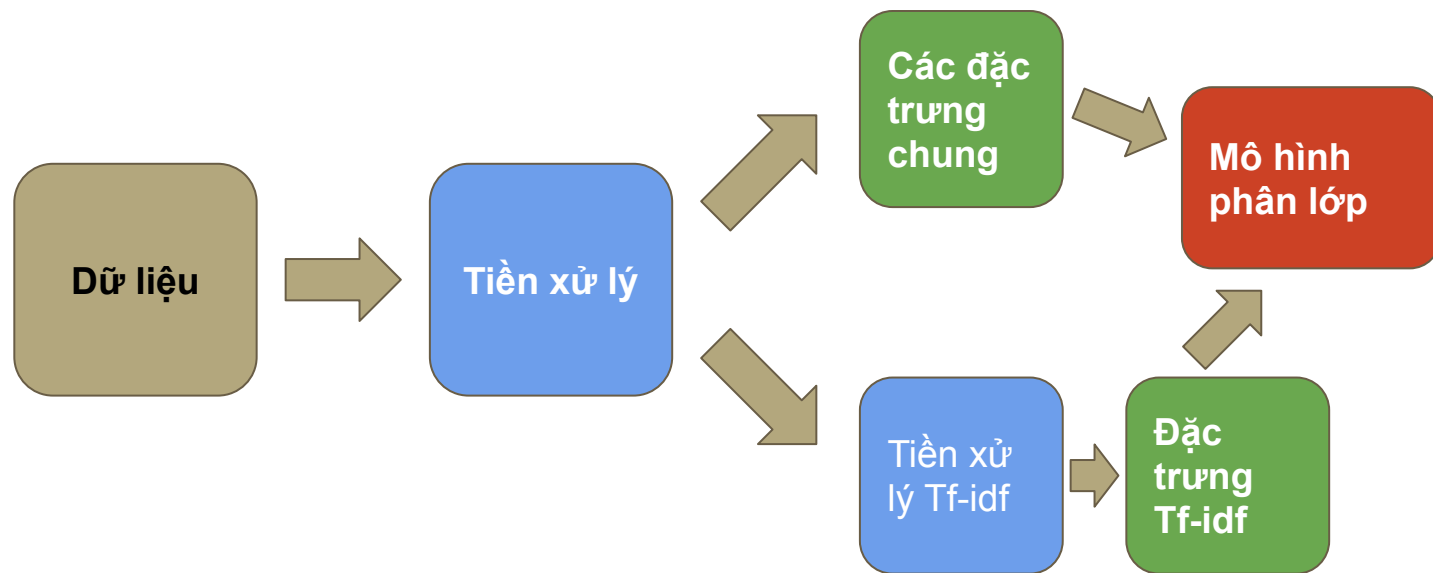
$$\mathbf{tf-idf(A, D_i, D) = tf(A, D_i) * idf(A, D)}$$

Trong đó:

**tf(A, D<sub>i</sub>)** = Số lần xuất hiện của **từ A** trong **document D<sub>i</sub>**

**idf(A, D)** =  $\ln(\text{Tổng số document trong } D / \text{Số document trong } D \text{ mà từ } A \text{ xuất hiện})$

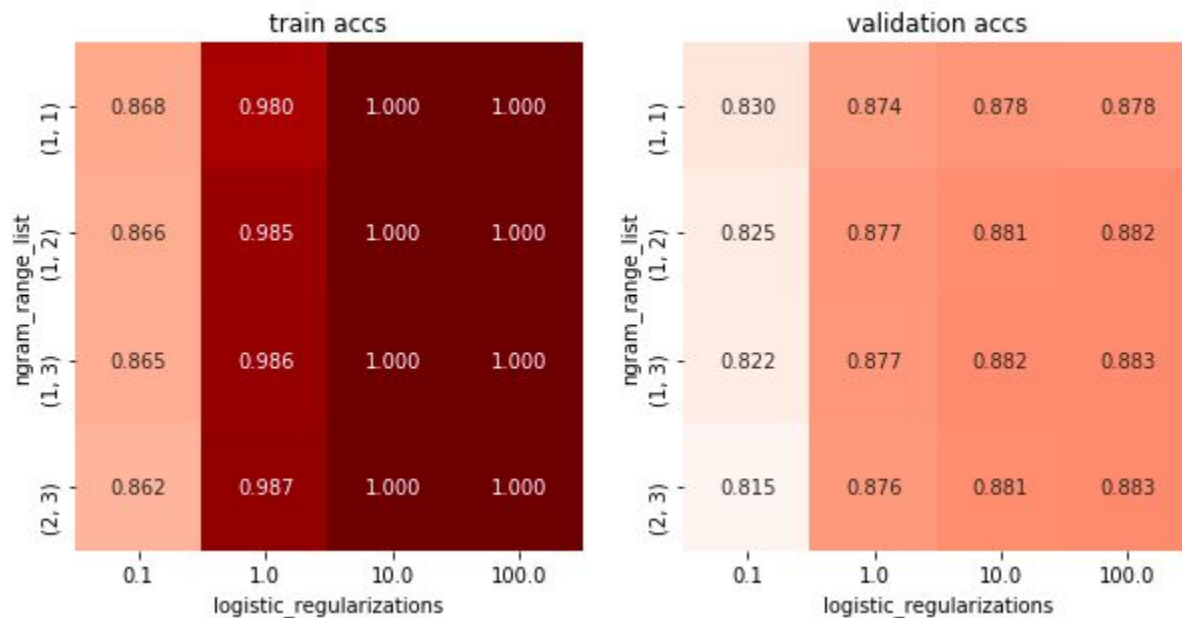
# Thiết kế mô hình



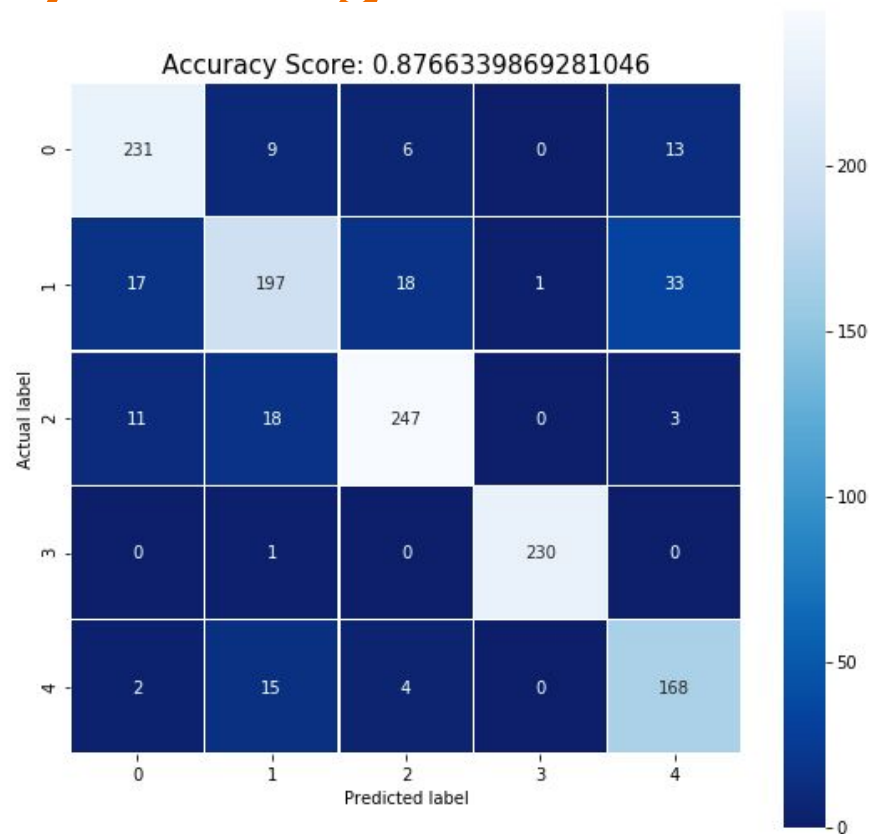
# Huấn luyện, đánh giá và kiểm thử

- 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