TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KÌ MÔN DỰ ÁN CÔNG NGHỆ THÔNG TIN

TÌM HIỂU VÀ ỨNG DỤNG ELASTICSEARCH CHO TÌM KIẾM TIN TỨC VỀ PHÁP LUẬT

Người hướng dẫn: **TS. LÊ ANH CƯỜNG**
Người thực hiện: **VINH LINH – 52000910**
VÕ THỊ QUẾ CHI – 52000741
Lớp **: 20050301**
Khoá **: 24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KÌ MÔN DỰ ÁN CÔNG NGHỆ THÔNG TIN

TÌM HIỂU VÀ ỨNG DỤNG ELASTICSEARCH CHO TÌM KIẾM TIN TỨC VỀ PHÁP LUẬT

Người hướng dẫn: **TS. LÊ ANH CƯỜNG**
Người thực hiện: **VINH LINH**
VÕ THỊ QUẾ CHI
Lớp **: 20050301**
Khoá **: 24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Để hoàn thành đồ án này chung em xin gửi đến thầy LÊ ANH CƯỜNG lời cảm ơn chân thành và sâu sắc nhât. Thầy đã tận tình hướng dẫn, giúp đỡ chúng em hoàn thành đề tài này. Với điều kiện thời gian cũng như kinh nghiệm còn hạn chế của sinh viê, bài báo cáo này không thể tránh khỏi được những thiếu sót. Chúng em rất mong nhận được sự giúp đỡ, đóng góp ý kiến từ thầy để chúng em nắm rõ hơn về kiến thức của môn áp dụng vào thực tế.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của chúng em và được sự hướng dẫn của TS Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Vinh Linh

Võ Thị Quế Chi

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

ERP Enterprise Resource Planning

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

CHƯƠNG 1 – GIỚI THIỆU

1.1 Vấn đề nghiên cứu và ý nghĩa

Trong thời đại số ngày nay, sự phổ biến của thông tin trên internet đã tạo nên một thách thức lớn đối với việc tìm kiếm và truy cập thông tin đáng tin cậy. Trong bối cảnh pháp luật, nhu cầu tìm kiếm thông tin pháp lý chính xác và nhanh chóng ngày càng trở nên quan trọng. Vì nhận thức về vấn đề này đã thúc đẩy đã thúc đẩy chúng em đến với tìm hiểu và ứng dụng Elasticsearch cho việc xây dựng một hệ thống tìm kiếm tin tức pháp luật trên website.

Elasticsearch là một công cụ mạnh mẽ về tìm kiếm và phân tích dữ liệu, đã chứng minh sức mạnh của mình trong việc xử lý lượng thông tin lớn và đa dạng. Qua dự án này, chúng em sẽ khám phá cách Elasticsearch có thể được tích hợp vào hệ thống của chúng em để cải thiện hiệu suất tìm kiếm, đồng thời tạo ra một trải nghiệm người dùng thuận lợi và hiệu quả.

Dự án này không chỉ mang lại kiến thức về Elasticsearch mà còn góp phần giải quyết một vấn đề thiết thực trong lĩnh vực pháp luật và tin tức. Chúng em hy vọng kết quả của dự án này sẽ là một bước tiến tốt trong việc cải thiện khả năng tìm kiếm thông tin pháp lý trực tuyến, hỗ trợ người dùng trong việc nắm bắt nhanh chóng và chính xác những thay đổi trong lĩnh vực pháp luật đang diễn ra không ngừng.

1.2 Tổng quan về các nghiên cứu liên quan và các công trình đã có:

- Một số nghiên cứu liên quan đến Elastic search mà chúng em đã tham khảo:
  - Tìm hiểu về Elastic Search. Ứng dụng của Elastic Search của Cloud Geeks Vietnam : Bài viết này giới thiệu về Elasticsearch, cách cài đặt và sử dụng Elasticsearch trong các ứng dụng tìm kiếm.
  - Lập trình chức năng tìm kiếm trong PHP với Elasticsearch của XuanThuLab: Bài viết này giới thiệu cách sử dụng Elasticsearch để tìm kiếm trong PHP.
  - Scripted Field trong Elasticsearch của tác giả Nguyễn tấn Nam: nghiên cứu kiến thức về Scripted Field và có thể sử dụng tính năng này để tăng cường khả năng tìm kiếm và phân tích dữ liệu trong Elasticsearch.
- Một số công ty lớn đã sử dụng Elasticsearch trong các sản phẩm của họ để tìm kiếm và phân tích dữ liệu trên trang web
  - Wikipedia: Wikipedia, trang web thông tin hàng đầu thế giới, đã tích hợp Elasticsearch để cải thiện trải nghiệm tìm kiếm cho người đọc. Việc này giúp người dùng nhanh chóng tìm thấy thông tin trên các trang wiki với hiệu suất cao.

- GitHub: GitHub, nền tảng lưu trữ mã nguồn hàng đầu, sử dụng Elasticsearch để hỗ trợ tính năng tìm kiếm và lọc thông tin trong các dự án mã nguồn mở. Điều này giúp các nhà phát triển dễ dàng tìm thấy mã nguồn, vấn đề, và các tài liệu liên quan.
- Stack Overflow: Stack Overflow, cộng đồng chia sẻ kiến thức lập trình, cũng là một ví dụ về việc sử dụng Elasticsearch. Họ tích hợp công cụ này để tối ưu hóa việc tìm kiếm câu hỏi và câu trả lời, giúp người dùng dễ dàng tìm thấy thông tin cần thiết.
- Netflix: Netflix, dịch vụ phát sóng video trực tuyến, đã áp dụng Elasticsearch trong hệ thống tìm kiếm để cung cấp trải nghiệm tìm kiếm linh hoạt và chính xác cho người dùng khi tìm kiếm nội dung trên nền tảng của họ.
- LinkedIn: LinkedIn, mạng xã hội chuyên nghiệp, cũng đã sử dụng Elasticsearch để cải thiện khả năng tìm kiếm thông tin người dùng, việc tìm kiếm công việc và nội dung chia sẻ trên nền tảng của họ.

1.3 Mục tiêu và phương pháp nghiên cứu:

1.3.1 Mục tiêu nghiên cứu:

Mục tiêu của đề tài này là nghiên cứu và phát triển Elasticsearch xây dựng hệ thống tìm kiếm thông tin để cải thiện khả năng tìm kiếm và phân tích văn bản ứng dụng vào website tin tức pháp luật. Nâng cao trải nghiệm người dùng, tăng cường khả năng phân tích và trích xuất thông tin, để người dùng có thể nắm bắt thông tin một cách nhanh chóng.

1.3.2 Phương pháp nghiên cứu:

- Hiểu rõ về Elasticsearch: Nắm vững kiến thức về Elasticsearch, bao gồm cách nó hoạt động, cấu trúc dữ liệu, và các tính năng quan trọng. Đánh giá khả năng mở rộng và hiệu suất của Elasticsearch để đảm bảo sự phù hợp với nhu cầu tìm kiếm tin tức pháp luật.

- Tìm hiểu về dữ liệu pháp luật: Nghiên cứu sâu rộng về cấu trúc dữ liệu và đặc điểm của tin tức pháp luật để hiểu rõ nhu cầu tìm kiếm của người dùng. Xác định các đặc điểm đặc biệt của thông tin pháp luật, bao gồm thuật ngữ chuyên ngành và quan hệ giữa các khái niệm.

- Thiết kế mô hình tìm kiếm hiệu quả: Phát triển một mô hình tìm kiếm sử dụng Elasticsearch được tối ưu hóa cho việc truy vấn thông tin pháp luật. Tích hợp các tính năng như lọc thông tin, sắp xếp theo độ ưu tiên, và gợi ý tìm kiếm để cải thiện trải nghiệm người dùng.

- Kiểm thử và tối ưu hóa hiệu suất: Tiến hành kiểm thử thực nghiệm để đánh giá hiệu suất của hệ thống tìm kiếm trong môi trường thực tế. Tối ưu hóa cấu hình Elasticsearch để đảm bảo độ nhanh chóng và đáng tin cậy trong việc truy vấn tin tức pháp luật.

- Thiết kế giao diện: Phát triển giao diện người dùng thân thiện và dễ sử dụng, tương thích với nhu cầu của người dùng cuối. Đảm bảo tính tương tác cao và trải nghiệm người dùng tích cực.

1.4 Đối tượng nghiên cứu:

- Công cụ Elasticsearch: Sử dụng phương pháp nghiên cứu thực nghiệm để xây dựng, triển khai và đánh giá hiệu suất của hệ thống tìm kiếm sử dụng Elasticsearch.

Áp dụng phương pháp thực hành để kết hợp mô hình RAG vào hệ thống tìm kiếm.

- Mô hình nghiên cứu: Áp dụng mô hình RAG (Retrieval-Augmented Generation) để cải thiện quá trình tìm kiếm thông tin pháp luật và sinh nội dung mô tả chi tiết. Nghiên cứu và xác định cách tích hợp mô hình RAG vào cấu trúc hệ thống Elasticsearch để tối ưu hóa trải nghiệm người dùng.

- Dữ liệu nghiên cứu: Sử dụng dữ liệu thực tế về tin tức pháp luật, bao gồm các văn bản pháp lý, thông tin về vụ án, và bài viết tin tức pháp luật. Dữ liệu này sẽ được sử dụng để xây dựng và kiểm thử hệ thống Elasticsearch cũng như để đào tạo mô hình RAG.

- Nghiệp vụ: Tập trung vào lĩnh vực pháp luật, nghiên cứu sẽ tìm hiểu rõ về yêu cầu và nhu cầu cụ thể của người dùng trong lĩnh vực này.

CHƯƠNG II – KIẾN THỨC NỀN TẢNG

2.1 Giới thiệu về Elasticsearch

Elasticsearch nói đơn giản là một cơ sở dữ liệu dùng để lưu trữ, tìm kiếm và query.

Elasticsearch là một công cụ mã nguồn mở tìm kiếm toàn văn bản và có khả năng phân tích mở rộng. Nó cho phép người dùng lưu trữ, tìm kiếm và phân tích khối lượng lớn dữ liệu một cách nhanh chóng và gần thời gian thực. Nó thường được sử dụng như các công cụ, công nghệ cơ bản mà quyền hạn các ứng dụng có tính năng tìm kiếm phức tạp và yêu cầu.

Elasticsearch là một công cụ dựa trên phần mềm Lucene. Nó cung cấp một bộ máy tìm kiếm dạng phân tán, có đầy đủ công cụ với một giao diện web HTTP có hỗ trợ dữ liệu JSON. Elasticsearch được phát triển bằng Java và được phát hành dạng mã nguồn mở theo giấy phép Apache.

- Các tính năng chính của elasticsearch:
  - Phân tích thời gian
  - Phân phối lưu trữ tập tin theo thời gian thực và từng lĩnh vực được lập chỉ mục.
  - Tài liệu định hướng, tất cả các đối tượng là tất cả các tài liệu.
  - Có tính sẵn sàng cao, dễ dàng mở rộng, hỗ trợ cluster, phân mảnh và nhân rộng (Shards và bản sao).
  - Giao diện thân thiện, hỗ trợ cho JSON.
  - Tìm Kiếm Toàn Văn Bản: Elasticsearch cung cấp một công cụ tìm kiếm mạnh mẽ cho văn bản đầy đủ và linh hoạt, bao gồm cả tìm kiếm phần vănbản, sắp xếp và lọc kết quả.
  - Phân Tán và Mở Rộng: Khả năng mở rộng của Elasticsearch cho phép nómở rộng dữ liệu và số lượng người dùng mà không làm giảm hiệu suất.
  - Tìm Kiếm Đồng Nhất (Consistency): Elasticsearch hỗ trợ các mô hình tìmkiếm đồng nhất thông qua các cấu hình nhất quán (consistency levels).
  - Strong Consistency (Đồng Nhất Mạnh): Điều này đảm bảo rằng mọi đọc và ghi sẽ thấy dữ liệu ở một trạng thái đồng nhất trên toàn bộ hệ thống. Tính đồng nhất mạnh là đảm bảo rằng mọi replica sẽ đồng bộ với nhau mộtcách ngay lập tức.

- Eventual Consistency (Đồng Nhất Cuối Cùng): Điều này cho phép một số độ trễ giữa khi một giá trị được ghi và khi nó trở nên đồng nhất trên toàn bộ hệ thống. Eventual consistency không đảm bảo ngay lập tức, nhưng nó hứa rằng cuối cùng, tất cả các bản sao sẽ đồng bộ.
- Causal Consistency (Đồng Nhất Nhân Quả): Điều này giữ cho các sự kiện xảy ra có liên quan với nhau sẽ được duy trì trong một thứ tự nhất định trên tất cả các node. Nó đảm bảo rằng nếu một sự kiện A gây ra sự kiện B, thì tất cả các node sẽ thấy B sau A.)
- Chức Năng Query DSL (Domain-Specific Language): Elasticsearch sử dụng một ngôn ngữ truy vấn có cú pháp đặc biệt, được gọi là Query DSL, cho phép người dùng xác định các truy vấn phức tạp.
- Xử Lý Nâng Cao Dữ Liệu: Có khả năng xử lý dữ liệu phức tạp, bao gồm cả các loại dữ liệu khác nhau như văn bản, số, địa lý và hình ảnh.
- Giao Diện RESTful API: Elasticsearch cung cấp một giao diện APIRESTful, làm cho việc tương tác và tích hợp với các ứng dụng web và dịchvụ trở nên dễ dàng.

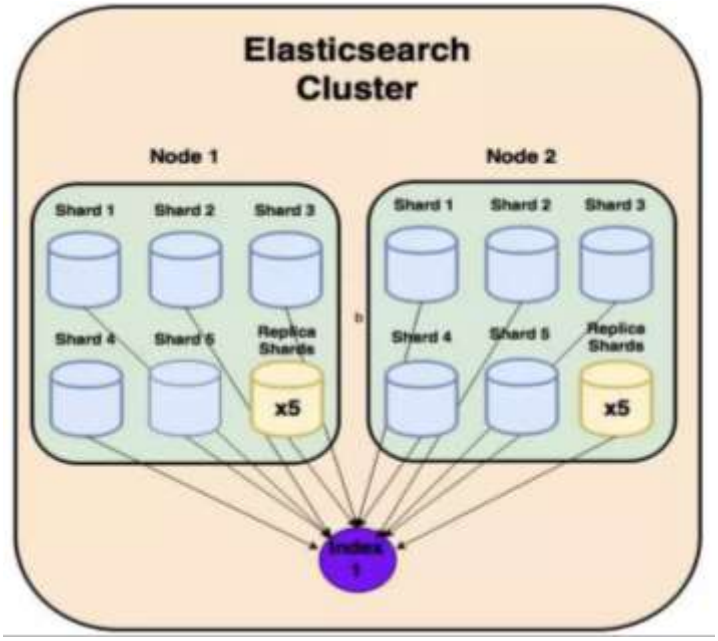
➢ Các thành phần của Elasticsearch:

- Index: là một tập hợp các document có điểm tương tự nhau. Có 2 loại:
  + Forward Index: Đánh index theo nội dung, page -> words.
  + Inverted Index: Đánh index theo keyword: word -> pages.
- Document: một JSON object với một số dữ liệu. Đây là đơn vị dữ liệu cơ bản trong Elasticsearch.
  + JSON (JavaScript Object Notation) là một kiểu định dangj dữ liệu tuân theo một quy luật nhất định mà hầu hết các ngôn ngữ lập trình hiện nay đều có thể đọc được.
- Shard: Tập con các document của một index. Một index có thể có nhiều shard. Có 2 loại:

+ Primary Shard: Là nơi lưu trữ dữ liệu và đánh index. Sau khi đánh xong dữ liệu sẽ được vận chuyển tới các Replica Shard.

+ Replica Shard: Là nơi lưu trữ dữ liệu nhân bản của Primary Shard, bảo vệ duẽ liệu toàn vẹn khi có sự cố.

- Node: Nơi lưu trữ dữ liệu, tham gia vào việc đánh mục lục của cluster cũng như thực hiện việc tìm kiếm. Mỗi node được định danh bằng một unique name.
- Cluster: Tập hợp các node chứa tất cả các dữ liệu. Mỗi cluster được định danh bằng một unique name. Mỗi cluster có một node chính (master) được lựa chọn tự động và có thể thay thế khi gặp sự cố.
  - Cách lưu cấu trúc cơ bản của 1 cluster trong Elasticsearch:



Hình 1: Cấu trúc của 1 cluster trong Elasticsearch.

2.2 Mô hình RAG:

Mô hình RAG (Retrieval-Augmented Generation) là một kiến trúc học máy đa nhiệm sử dụng kết hợp giữa khả năng truy vấn và sinh nội dung. Mô hình này được thiết kế để giải quyết các vấn đề liên quan đến tìm kiếm thông tin và sinh nội dung mô tả chi tiết trong một hệ thống duyệt web hoặc ứng dụng tìm kiếm.

- Kiến trúc RAG:

  + Truy xuất thông tin (Retrieval): Tìm kiếm các tài liệu có liên quan đến đầu vào của người dùng trong một kho dữ liệu lớn. Thường sử dụng kỹ thuật nhúng vector (vector embedding) và cơ sở dữ liệu vector để truy vấn nhanh chóng và hiệu quả.

  + Tạo sinh văn bản (Generation): Dựa trên các tài liệu được truy xuất, mô hình sinh văn bản tạo ra đoạn văn bản mới phù hợp với yêu cầu của người dùng. Cácmô hình ngôn ngữ lớn (Large Language Models - LLM) như BERT, GPT-3 thường được sử dụng trong giai đoạn này.

- Khả năng sinh nội dung tự nhiên:

  + Một trong những điểm mạnh của RAG là khả năng sinh nội dung tự nhiên và phức tạp. Mô hình này có thể tạo ra các đoạn văn bản có ý nghĩa và thông tin chi tiết, giúp người dùng hiểu rõ hơn về kết quả tìm kiếm.

- Chuỗi mô hình:

  + Mô hình RAG thường được xây dựng theo chuỗi, có nghĩa là mô-đun truy vấn và mô-đun sinh nội dung hoạt động lần lượt. Điều này giúp tối ưu hóa quá trình và tạo ra kết quả tốt nhất cho người dùng.

- Tích hợp mô hình ngôn ngữ tự nhiên:

  + Mô hình RAG thường tích hợp với các mô hình ngôn ngữ tự nhiên như GPT (Generative Pre-trained Transformer) để cải thiện khả năng sinh nội dung và hiểu ngôn ngữ.

- Học chia sẻ kiến thức:

  + Mô hình RAG có khả năng học chia sẻ kiến thức giữa mô-đun truy vấn và mô-đun sinh nội dung, giúp chúng hiểu rõ và tận dụng thông tin một cách hiệu quả.

- Ứng dụng trong tìm kiếm thông tin:

+ Mô hình RAG thường được ứng dụng trong các hệ thống tìm kiếm thông tin để cải thiện trải nghiệm người dùng. Nó có thể được sử dụng trong các ứng dụng như duyệt web, ứng dụng di động, và các dự án tìm kiếm thông tin phức tạp.

- Tối ưu hóa hiệu suất và độ chính xác:

  + Mô hình RAG được thiết kế để cân bằng giữa hiệu suất và độ chính xác. Nó cung cấp kết quả tìm kiếm nhanh chóng và đồng thời đảm bảo độ chính xác và chất lượng của thông tin sinh nội dung.

❖ Quá trình hoạt động của RAG với dạng cơ sở dữ liệu vector

- Lưu trữ dữ liệu : Khi tài liệu được nhập vào, một mô hình nhúng sẽ biến văn bản thành vectơ nhiều chiều. Mỗi vectơ thể hiện ý nghĩa ngữ nghĩa của tài liệu dưới dạng số. Các vectơ này sau đó được lưu trữ trong cơ sở dữ liệu vectơ.
- Lập chỉ mục : Để tạo điều kiện truy xuất nhanh, cơ sở dữ liệu xây dựng chỉmục trên các vectơ này bằng các thuật toán phù hợp với không gian nhiềuchiều, chẳng hạn như Chỉ mục tệp đảo ngược (IVF) hoặc Thế giới nhỏ điềuhướng phân cấp (HNSW). Việc lựa chọn loại chỉ mục sẽ cân bằng giữa tốc độ và độ chính xác của tìm kiếm.
- Tìm kiếm tương tự : Khi một truy vấn được thực hiện, nó cũng được chuyển đổi thành một vectơ bằng cách sử dụng cùng một mô hình nhúng. Cơ sở dữ liệu vectơ sau đó sử dụng chỉ mục để nhanh chóng tìm thấy các vectơ giống nhất với vectơ truy vấn. Độ tương tự được xác định bằng các số liệu khoảng cách như khoảng cách Euclide hoặc độ tương tự cosin.

➢ Nhìn chung, mô hình RAG đại diện cho một tiến bộ quan trọng trong lĩnh vực tìm kiếm thông tin và sinh nội dung tự nhiên, mang lại sự linh hoạt và khả năng hiểu biết cao. Sự kết hợp giữa khả năng truy vấn và sinh nội dung mô tả chi tiết làm

cho RAG trở thành một công cụ mạnh mẽ trong các dự án tìm kiếm thông tin và môi trường web.

CHƯƠNG 3 – PHƯƠNG PHÁP ĐỀ XUẤT

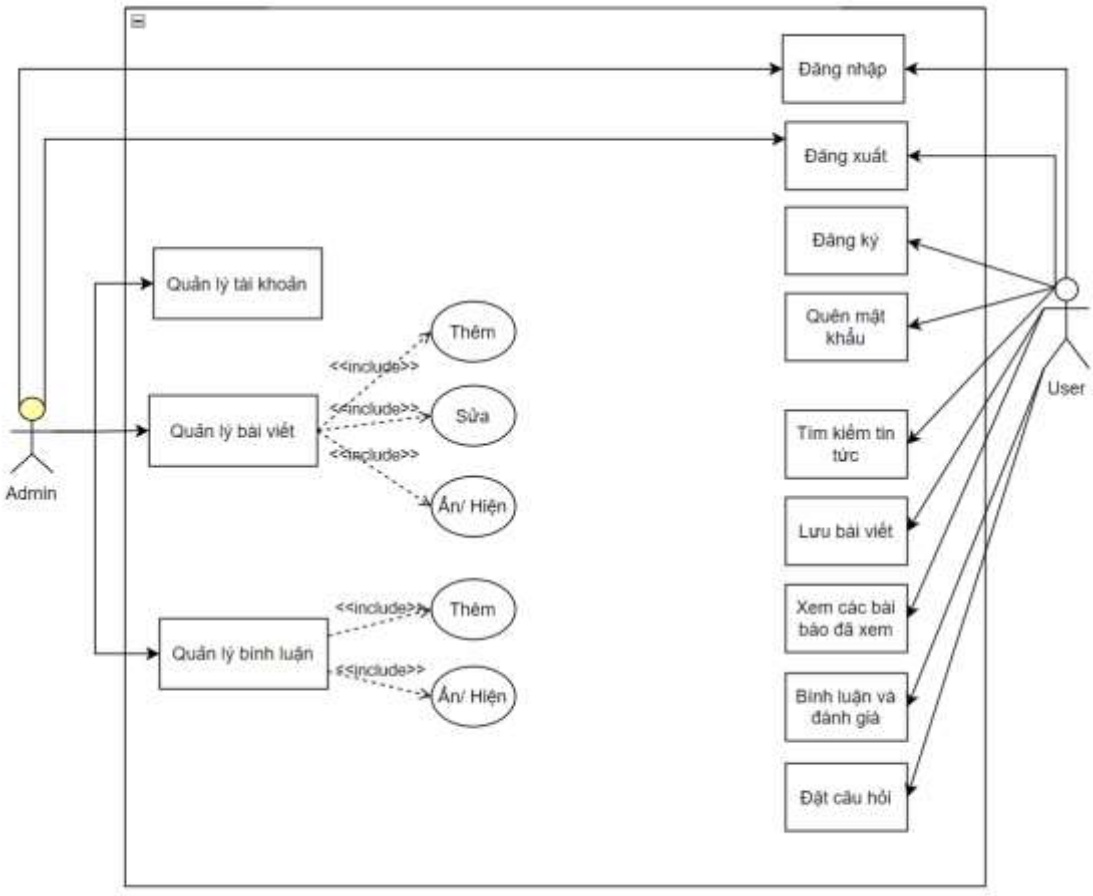
3.1 Đặc tả hệ thống:

Website tìm kiếm tin tức pháp luật phục vụ cho những người dùng có nhu cầu đọc tin tức về pháp luật một cách nhanh nhất và chính xác.

Về phía người dùng có thể tìm kiếm đến các bài báo về pháp luật mà mình muốn đọc, đăng ký, đăng nhập, lưu bài báo, xem các bài báo đã đọc, bình luận và đánh giá các bài báo, hoặc đặt ra câu hỏi về pháp luật Việt Nam.

Về phía Admin có thể quản lý các bài báo (thêm, sửa, ẩn,...), quản lý bình luận, quản lý tài khoản.

3.2 Biểu đồ Usecase tổng quát:



Hình 2: Usecase tổng quát

3.3 Đặc tả Use case:

3.3.1 Use case đăng ký:

Use case	Đăng ký	
Ngữ cảnh	Tác nhân muốn đăng ký tài khoản	
Sự kiện kích hoạt	Tác nhân ấn nút Đăng ký.	
Mô tả	Tác nhân đăng ký để có thể đăng nhập vào website.	
Actors	Người dùng	
Điều kiện tiên quyết	Tác nhân truy cập vào website.	
Kết quả	Nếu usecase này thực hiện thành công, tác nhân sẽ có 1 tài khoản gồm tên đăng nhập và mật khẩu để đăng nhập vào website.	
Luồng sự kiện	Tác nhân	Hệ thống
	1. Tác nhân đang ở trang chủ để đăng kí vào hệ thống.	1.1 Hệ thống yêu cầu tác nhân nhập các thông tin cá nhân: Tên đăng nhập, Mật khẩu, xác nhận lại Mật khẩu,...

	2. Tác nhân nhập thông tin cá nhân: Tên đăng nhập, Mật khẩu,...và nhấn nút Đăng ký.	2.1 Hệ thống kiểm tra thông tin điền vào có hợp lệ hay không. 2.2 Hệ thống hiển thị Đăng ký thành công. 2.3 Use case kết thúc.
Ngoại lệ	Nếu không điền đủ các trường bắt buộc thì hệ thống sẽ thông báo Đăng ký thất bại và quay lại cho Tác nhân nhập lại thông tin để đăng ký.	

3.3.2 Use case đăng nhập:

Use case	Đăng nhập
Ngữ cảnh	Tác nhân muốn đăng nhập vào website.
Sự kiện kích hoạt	Tác nhân nhấn nút Đăng nhập.
Mô tả	Tác nhân đăng nhập vào website để thực hiện các thao tác: bình luận, đánh giá, xem các bài viết đã lưu,…
Actors	Người dùng, Admin
Điều kiện tiên quyết	Tác nhân đã có tài khoản, mật khẩu để truy cập vào website

Kết quả	Nếu use case này được thực hiện thành công, tác nhân sẽ đăng nhập vào hệ thống đã được cấp quyền. Ngược lại thì tác nhân không thể truy cập vào hệ thống phân quyền mà chỉ ở trạng thái đăng nhập, đăng ký.	
Luồng sự kiện	Tác nhân	Hệ thống
	1. Tác nhân đã truy cập vào trang chủ để đăng nhập vào hệ thống. 2. Tác nhân nhập Tên đăng nhập và Mật khẩu. Sau đó, nhấn nút Đăng nhập.	1.1 Hệ thống yêu cầu tác nhân nhập Tên đăng nhập và Mật khẩu đã cung cấp. 2.1 Hệ thống kiểm tra thông tin điền vào có trùng khớp với tài khoản đã cung cấp hay không. 2.2 Hệ thống hiển thị trạng thái Đăng nhập. 2.2.1 Nếu thông tin trùng khớp thì tác nhân được phép truy cập và trang đã được phân quyền trước đó. 2.2.2 Ngược lại thì tác nhân không thể truy cập vào hệ thống và vẫn ở trang đăng nhập, đăng ký. 2.3 Use case kết thúc.

Ngoại lệ	Nếu không điền đủ tên đăng nhập/mật khẩu thì hệ thống sẽ thông báo Đăng nhập thất bại và quay lại cho Tác nhân nhập lại thông tin tài khoản.

3.3.3 Use case đăng xuất:

Use case	Đăng xuất
Ngữ cảnh	Tác nhân muốn đăng xuất tài khoản khỏi hệ thống
Sự kiện kích hoạt	Tác nhân nhấn nút Đăng xuất
Mô tả	Tác nhân thực hiện thao tác đăng xuất trên thanh công cụ của hệ thống. Sau đó, giao diện quay về trang chủ đầu tiên khi tác nhân đăng nhập/ đăng ký thì tài khoản đã đăng xuất thành công
Actors	Người dùng, Admin
Điều kiện tiên quyết	Tác nhân đã đăng nhập thành công vào hệ thống
Kết quả	Đăng xuất tài khoản khỏi hệ thống thành công, giao diện quay về trạng thái đăng nhập/ đăng ký ban đầu

Luồng sự kiện	Tác nhân	Hệ thống
	1. Tác nhân nhấn nút đăng xuất trên thanh công cụ 2. Tác nhân tải lại trang và kiểm tra kĩ đã đăng xuất thành công hay chưa	1.1 Tài khoản của tác nhân thoát khỏi hệ thống và quay về trạng thái đăng nhập/ đăng ký ban đầu. 2.1 Tài khoản đã đăng xuất hoàn toàn khỏi hệ thống. 2.2 Use case kết thúc
Ngoại lệ	Hệ thống bị lỗi không thể đăng xuất khỏi hệ thống.	

3.3.4 Use case tìm kiếm bài báo:

Use case	Tìm kiếm thông tin
Ngữ cảnh	Tác nhân tìm kiếm và xem dữ liệu theo yêu cầu mong muốn
Sự kiện kích hoạt	Tác nhân muốn tìm kiếm trên thông tin, dữ liệu

Mô tả	Tác nhân muốn thực hiện tìm kiếm thông tin, dữ liệu: tìm kiếm tin tức trên thanh tìm kiếm.	
Actors	Người dùng	
Điều kiện tiên quyết	Tác nhân đã vào website.	
Kết quả	Truy cập và xem được thông tin, dữ liệu cần tìm.	
Luồng sự kiện	Tác nhân	Hệ thống
	1. Tác nhân chọn mục tìm kiếm. 2. Tác nhân nhập thông tin, dữ liệu ngắn gọn cần tìm kiếm 3. Tác nhân chọn thông tin, dữ liệu đúng cần tìm 4. Tác nhân xem thông tin, dữ liệu cần tìm	1.1 Hệ thống hiển thị chỗ để tác nhân nhập thông tin, dữ liệu cần tìm kiếm 2.1 Hệ thống hiển thị tất cả các thông tin, dữ liệu liên quan lên màn hình. 3.1 Hệ thống hiển thị đúng thông tin, dữ liệu tác nhân yêu cầu 4.1 Use case kết thúc

Ngoại lệ	Nếu không tìm thấy thông tin, dữ liệu, hệ thống sẽ thông báo dữ liệu, thông tin không tồn tại

3.3.5 Use case xem chi tiết bài báo:

Use case	Xem chi tiết bài báo	
Ngữ cảnh	Tác nhân muốn xem thông tin chi tiết tin tức	
Sự kiện kích hoạt	Tác nhân muốn xem thông tin chi tiết tin tức đã được đăng trên website.	
Mô tả	Use case này cho phép tác nhân xem toàn bộ thông tin về bài báo đã chọn	
Actors	Người dùng	
Điều kiện tiên quyết	Tác nhân đã vào website	
Kết quả	Tác nhân có thể xem được toàn bộ thông tin cần tìm về bài báo đã đăng	
Luồng sự kiện	Tác nhân	Hệ thống

	1. Tác nhân vào hệ thống tìm kiếm tin tức pháp luật muốn xem 2. Tác nhân chọn chọn chọn bài bái cần đọc	1.1 Hệ thống phản hồi và hiển thị tất cả các bài báo liên quan đến từ khóa cần tìm 2.1 Hệ thống hiển thị toàn bộ thông tin mới nhất về bài báo được chọn 2.2 Use case kết thúc
Ngoại lệ	Nếu kết nối mạng bị lỗi có thể xảy ra trường hợp lỗi.	

3.3.6 Use case lưu bài báo:

Use case	Lưu bài báo
Ngữ cảnh	Tác nhân đánh dấu lại những bài báo hay hoặc cần đọc lại
Sự kiện kích hoạt	Tác nhân muốn đánh dấu và lưu giữ lại các bài báo
Mô tả	Use case cho phép tác nhân đánh dấu và lưu giữ bài báo có thể sẽ đọc lại và hệ thống sẽ đề xuất các bài báo tương tự
Actors	Người dùng
Điều kiện tiên quyết	Tác nhân đã đăng nhập vào hệ thống đã cấp quyền

Kết quả	Thông tin về bài báo được lưu giữ lại và người dùng sẽ dễ dàng xem lại	
Luồng sự kiện	Tác nhân	Hệ thống
	1.Tác nhân thực hiện chọn nút đánh dấu hoặc mục theo dõi trên 2. Tác nhân vào mục đã lưu hoặc đang theo dõi để xem lại thông tin.	1.1 Hệ thống hiển thị thông báo đã lưu hoặc đang theo dõi 2.1 Hệ thống hiển thị toàn bộ các thông tin đã được lưu trữ. 2.2 Use case kết thúc
Ngoại lệ	Nếu hệ thống kết nối với đường truyền internet bị lỗi thì không thể đánh dấu, lưu giữ thông tin, dữ liệu vào hệ thống được	

3.4 Sơ đồ lớp (Class Diagram):

3.5 Đặc tả cơ sở dữ liệu:

3.5.1 Thông tin và tài khoản của người dùng:

Tên cột dữ liệu	Kiểu dữ liệu
Mã tài khoản: STT (PK)	int
Tên: name_user	String
Giới tính: gender	String
Email:email_user	String
Mật khẩu:pass_user	String

3.5.2 Thông tin bài báo:

Tên cột dữ liệu	Kiểu dữ liệu
Mã bài báo: STT (PK)	int
Tiêu đề bài báo: title	String
Tác giả: author	String
Mô tả: description	String
Hình ảnh: img	String
Nội dung: text	String

3.6 Giới Thiệu về Dataset

CHƯƠNG 4 – THỰC NGHIỆM

4.1 Thu thập dữ liệu

- Giới Thiệu VIETNAMESE FAKE NEWS DATASET – VFND

VFND là bộ dataset về các tin tức giả bằng ngôn ngữ tiếng Việt được tập hợp trong khoảng thời gian từ 2017 đến 2019 (năm hoàn thành luận văn của các tác giả), các tin tức được đưa vào đây được phân loại thật giả dựa trên một số nguồn tin, thamchiếu chéo đến các nguồn tin được dẫn hoặc được phân loại bởi cộng đồng.

❖ Dữ liệu file CSV

Có 2 label để phân loại là: Fake (tin tức giả) giá trị 1 và Real (tin tức thật) giá trị 0. Các file và ý nghĩa têncủa chúng:

+ vn_news_226_tlfr.csv: Chứa 226 record dữ liệu bao gồm 2 trường Text và Label. Text tổng hợp từ các tin tức giả và thật từ Báo chí, tin tức báo chí bao gồm phần tiêu đề và nội dung.

+ vn_news_223_tdlfr.csv: Chứa 223 record dữ liệu các bài báo và domain name của các trang đã đăng các bài báo đó. tdlfr là [text, domain, label, fake,real]

❖ Dữ liệu file CSV và Json

Tập dữ liệu gồm: Tập dữ liệu được phân loại dựa trên 2 nhãn Fake và Real

trong đó với mỗi nhãn tương ứng với một thư muc, trong từng thư muc sẽ được phân loại thành 2 nhóm: Article_Contents: tập hợp các bài báo đã được phân

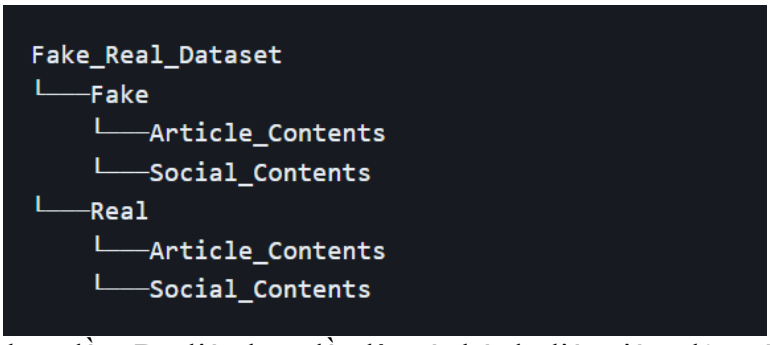
loại trong nhãn và Social_Contents: tập hợp các Facebook post đã được phânloại.

Cấu trúc của phần lưu trữ tin tức giả: {Tên file}: {Tiêu đề/Link} -

{Lý do phân loại}. Trong một số tin tức sẽ có thêm phầnrelevant_site"

VD : VFND_Ac_Fake_1 , VFDN_So_Fake_1

VD : VFND_Ac_Real_1, VFDN_So_Real_1



- Dữ liệu ban đầu: Dữ liệu ban đầu là một bộ dữ liệu riêng lẻ, một bài báo là một file json, tương ứng sẽ có 127 bài báo fake (127 file json) và 131 bài báo real (131 file json).

- Xử lý lại dữ liệu: Gọp hết thành một file json và chuyển sang file csv.

Gọp các file json:

```
, 'VFND_Ac_Real_80.json', 'VFND_Ac_Real_81.json', 'VFND_Ac_Real_82.json', 'VFND_Ac_Real_83.json'
, 'VFND_Ac_Real_84.json', 'VFND_Ac_Real_85.json', 'VFND_Ac_Real_86.json', 'VFND_Ac_Real_87.json'
, 'VFND_Ac_Real_88.json', 'VFND_Ac_Real_89.json', 'VFND_Ac_Real_90.json', 'VFND_Ac_Real_91.json'
, 'VFND_Ac_Real_92.json', 'VFND_Ac_Real_93.json', 'VFND_Ac_Real_94.json', 'VFND_Ac_Real_95.json'
, 'VFND_Ac_Real_96.json', 'VFND_Ac_Real_97.json', 'VFND_Ac_Real_98.json', 'VFND_Ac_Real_99.json'
, 'VFND_Ac_Real_100.json', 'VFND_Ac_Real_101.json', 'VFND_Ac_Real_102.json', 'VFND_Ac_Real_103.json'
, 'VFND_Ac_Real_104.json', 'VFND_Ac_Real_105.json', 'VFND_Ac_Real_106.json', 'VFND_Ac_Real_107.json'
, 'VFND_Ac_Real_108.json', 'VFND_Ac_Real_109.json', 'VFND_Ac_Real_110.json', 'VFND_Ac_Real_111.json'
, 'VFND_Ac_Real_112.json', 'VFND_Ac_Real_113.json', 'VFND_Ac_Real_114.json', 'VFND_Ac_Real_115.json'
, 'VFND_Ac_Real_116.json', 'VFND_Ac_Real_117.json', 'VFND_Ac_Real_118.json', 'VFND_Ac_Real_119.json'
, 'VFND_Ac_Real_120.json', 'VFND_Ac_Real_121.json', 'VFND_Ac_Real_122.json', 'VFND_Ac_Real_123.json'
, 'VFND_Ac_Real_124.json', 'VFND_Ac_Real_125.json', 'VFND_Ac_Real_126.json', 'VFND_Ac_Real_127.json']

# Biến lưu trữ nội dung ghép
combined_data = []

# Đọc từng file JSON và thêm nội dung vào biến combined_data
for file_path in file_paths:
    with open(file_path, 'r', encoding='utf-8') as json_file:
        data = json.load(json_file)
        combined_data.append(data)

# Ghi nội dung ghép vào một file JSON mới
output_file_path = 'dataset.json'
with open(output_file_path, 'w', encoding='utf-8') as output_file:
    json.dump(combined_data, output_file, ensure_ascii=False, indent=4)
```

TÀI LIỆU THAM KHẢO