BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC NHA TRANG  
KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN TỐT NGHIỆP**

**ỨNG DỤNG THUẬT TOÁN PHÂN LOẠI VĂN BẢN TRONG XÂY DỰNG TRANG WEB TIN TỨC**

**Giảng viên hướng dẫn: Phạm Thị Kim Ngoan**

**Nguyễn Hải Triều**

**Sinh viên thực hiệ****n: Nguyễn Văn Hoàng**

**Mã số Sinh viên: 58131303**

**Lớp: 58-TH1**

Khánh Hoà - 2020

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC NHA TRANG  
KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN TỐT NGHIỆP**

**ỨNG DỤNG THUẬT TOÁN PHÂN LOẠI VĂN BẢN TRONG XÂY DỰNG TRANG WEB TIN TỨC**

**Giảng viên hướng dẫn: Phạm Thị Kim Ngoan**

**Nguyễn Hải Triều**

**Sinh viên thực hiện: Nguyễn Văn Hoàng**

**Mã số Sinh viên: 58131303**

**Lớp: 58-TH1**

Khánh Hoà - 2020

**TRƯỜNG ĐẠI HỌC NHA TRANG**

**Khoa/Viện**:…………………………….

**PHIẾU THEO DÕI TIẾN ĐỘ VÀ ĐÁNH GIÁ ĐỒ ÁN / KHÓA LUẬN TỐT NGHIỆP  
*(Dùng cho CBHD và nộp cùng báo cáo ĐA/KLTN của sinh viên)***

Tên đề tài:

Chuyên ngành:

Họ và tên sinh viên: …………………………………Mã sinh viên:

Người hướng dẫn (học hàm, học vị, họ và tên):

Cơ quan công tác:

**Phần đánh giá và cho điểm của người hướng dẫn (tính theo thang điểm 10)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tiêu chí đánh giá** | **Trọng số**  **(%)** | **Mô tả mức chất lượng** | | | | **Điểm** |
| **Giỏi** | **Khá** | **Đạt yêu cầu** | **Không đạt** |
| **9 - 10** | **7 - 8** | **5 - 6** | **< 5** |
| Xây dựng đề cương nghiên cứu | 10 |  |  |  |  |  |
| Tinh thần và thái độ làm việc | 10 |  |  |  |  |  |
| Kiến thức và kỹ năng làm việc | 10 |  |  |  |  |  |
| Nội dung và kết quả đạt được | 40 |  |  |  |  |  |
| Kỹ năng viết và trình bày báo cáo | 30 |  |  |  |  |  |
| **ĐIỂM TỔNG** | | | | | |  |

*Ghi chú: Điểm tổng làm tròn đến 1 số lẻ.*

**Nhận xét chung** (sau khi sinh viên hoàn thành ĐA/KLTN):

…………………………………………………………………………………………….…………………………………………………………………………………………….…………………………………………………………………………………………………………….………………………………………………………………………………………Đồng ý cho sinh viên: Được bảo vệ:  Không được bảo vệ: 

*Khánh Hòa, ngày…….tháng…….năm………* **Cán bộ hướng dẫn** *(Ký và ghi rõ họ tên)*

**TRƯỜNG ĐẠI HỌC NHA TRANG**

**Khoa/Viện:……………………………….**

**PHIẾU CHẤM ĐIỂM ĐỒ ÁN/ KHÓA LUẬN TỐT NGHIỆP**

***(Dành cho cán bộ chấm phản biện)***

Tên đề tài:

Chuyên ngành:

Họ và tên sinh viên: …………………………………Mã sinh viên:

Người phản biện (học hàm, học vị, họ và tên):

Cơ quan công tác:

**I. Phần đánh giá và cho điểm của người phản biện (tính theo thang điểm 10)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tiêu chí đánh giá** | **Trọng số**  **(%)** | **Mô tả mức chất lượng** | | | | **Điểm** |
| **Giỏi** | **Khá** | **Đạt yêu cầu** | **Không đạt** |
| **9 - 10** | **7 - 8** | **5 - 6** | **< 5** |
| Hình thức bản thuyết minh | 30 |  |  |  |  |  |
| Nội dung bản  thuyết minh | 30 |  |  |  |  |  |
| Kết quả nghiên cứu | 20 |  |  |  |  |  |
| Mức độ trích dẫn và sao chép | 20 |  |  |  |  |  |
| **ĐIỂM TỔNG** | | | | | |  |

*Ghi chú: Điểm tổng làm tròn đến 1 số lẻ.*

**Kết luận:**

Đồng ý cho sinh viên: Được bảo vệ:  Không được bảo vệ: 

*Khánh Hòa, ngày…….tháng………năm………..* **Cán bộ chấm phản biện** *(Ký và ghi rõ họ tên)*

**LỜI CAM ĐOAN**

Tôi xin cam đoan đề tài: “Ứng dụng thuật toán phân loại văn bản trong xây dựng trang web tin tức” là một công trình nghiên cứu độc lập không có sự sao chép của người khác. Đề tài là một sản phẩm mà tôi đã nỗ lực nghiên cứu trong quá trình học tập tại trường, tự nghiên cứu cũng như thực tập tại công ty 9thWonder. Trong quá trình viết bài có sự tham khảo một số tài liệu có nguồn gốc rõ ràng, dưới sự hướng dẫn của cô Phạm Thị Kim Ngoan và thầy Nguyễn Hải Triều – Giáo Viên Khoa Công nghệ Thông Tin trường Đại Học Nha Trang. Tôi xin cam đoan các nội dung và kết quả thực hiện trong đề tài là nghiên cứu của cá nhân, tôi xin chịu hoàn toàn trách nhiệm về các nội dung trong báo cáo đề tài của mình.

Khánh Hòa, ngày 10 tháng 7 năm 2020

Sinh viên thực hiện

*Nguyễn Văn Hoàng*

**LỜI CẢM ƠN**

Tôi xin chân thành cảm ơn Khoa *Công nghệ thông tin* Trường Đại học Nha Trang đã tạo điều kiện tốt cho tôi thực hiện tốt Đồ án tốt nghiệp này.

Ngoài ra, tôi xin chân thành cảm ơn quý thầy cô trong khoa *Công nghệ Thông tin* đã  
tận tình giảng dạy, trang bị cho tôi những kiến thức quý báu trong những năm học  
qua, giúp tôi có một nền tảng kiến thức vững chắc để hoàn thành đồ án này. Đặc biệt  
tôi xin chân thành cám ơn sự nhiệt tình hướng dẫn và đóng góp ý kiến của cô  
*Phạm Thị Kim Ngoan và thầy Nguyễn Hải Triều -* giảng viên khoa *Công nghệ Thông tin* đã giúp tôi hoàn thành tốt đồ án tốt nghiệp.

Mặc dù, tôi đã cố gắng hoàn thành đồ án thực tập trong phạm vi và khả năng  
cho phép nhưng chắc chắn vẫn không tránh khỏi những thiếu sót. Tôi kính  
mong nhận được sự thông cảm và tận tình đóng góp ý kiến của quý thầy cô và các bạn.

*Tôi xin chân thành cảm ơn!*

**TÓM TẮT KHOÁ LUẬN**

**…**

**MỤC LỤC**

[**LỜI MỞ ĐẦU 1**](#_Toc45002699)

[**CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI 2**](#_Toc45002700)

[**1.1 Lý do chọn đề tài 2**](#_Toc45002701)

[**1.2 Đối tượng và phạm vi nghiên cứu 3**](#_Toc45002702)

[**1.3 Mục đích nghiên cứu 3**](#_Toc45002703)

[**1.4 Nội dung thực hiện 4**](#_Toc45002704)

[**1.4.1 Về trang web: 4**](#_Toc45002705)

[**1.4.2 Về thuật toán: 4**](#_Toc45002706)

[**1.4.3 Về dữ liệu : 4**](#_Toc45002707)

[**1.5 Môi trường cài đặt và IDE thực hiện đề tài 4**](#_Toc45002708)

[**CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 5**](#_Toc45002709)

[**2.1 Khái quát về HTML 5**](#_Toc45002710)

[**2.2 Khái quát về CSS 5**](#_Toc45002711)

[**2.3 Khái quát về JavaScript 5**](#_Toc45002712)

[**2.4 Khái quát về MySQL 5**](#_Toc45002713)

[**2.5 Khái quát về PHP 6**](#_Toc45002714)

[**2.6 Framework Laravel 6**](#_Toc45002715)

[**2.6.1 Khái niệm 6**](#_Toc45002716)

[**2.6.2 Cấu hình yêu cầu khi cài đặt Laravel 6**](#_Toc45002717)

[**2.6.3 Cài đặt Laravel 7**](#_Toc45002718)

[**2.6.4 Cấu trúc thư mục của một project Laravel: 11**](#_Toc45002719)

[**2.7 NodeJS 13**](#_Toc45002720)

[**2.7.1 Khái niệm 13**](#_Toc45002721)

[**2.7.2 Những ứng dụng nên viết bằng Nodejs 13**](#_Toc45002722)

[**2.7.3 Những kết luận sai lầm về Nodejs 14**](#_Toc45002723)

[**2.7.4 Lý do nên sử dụng Nodejs 14**](#_Toc45002724)

[**2.7.5 Node Package Manager (NPM) 15**](#_Toc45002725)

[**2.7.6 Cài đặt NodeJS 15**](#_Toc45002726)

[**2.7.7 Sử dụng NodeJS cơ bản 15**](#_Toc45002727)

[**2.8 Các module xử lý dữ liệu và ứng dụng thuật toán Naive Bayes vào hệ thống 16**](#_Toc45002728)

[**2.9 Định lý Bayes 18**](#_Toc45002729)

[**2.10 Thuật toán phân lớp Naive Bayes Classification 21**](#_Toc45002730)

[**2.11 Laravel kết hợp NodeJS 21**](#_Toc45002731)

[**CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG 23**](#_Toc45002732)

[**3.1 Phân tích và thiết kế hệ thống WebSite 23**](#_Toc45002733)

[**3.1.1 Định nghĩa bài toán 23**](#_Toc45002734)

[**3.1.2 Sơ đồ phân ra chức năng 23**](#_Toc45002735)

[**3.1.3 Cơ sở dữ liệu 29**](#_Toc45002736)

[**3.2 Ứng dụng thuật toán phân loại Naive Bayes Classifier 32**](#_Toc45002737)

[**3.2.1 Tiến hành thu thập dữ liệu: 33**](#_Toc45002738)

[**3.2.2 Phân tích và tiền xử lý dữ liệu trong hệ thống: 36**](#_Toc45002739)

[**3.2.3 Tiến hành quá trình phân loại dữ liệu: 37**](#_Toc45002740)

[**3.3.4 Kết quả: 38**](#_Toc45002741)

[**3.2.5 Đánh giá: 38**](#_Toc45002742)

[**CHƯƠNG 4. KẾT QUẢ THỰC HIỆN ĐỀ TÀI 40**](#_Toc45002743)

[**4.1 Giao diện người dùng 40**](#_Toc45002744)

[**4.2 Giao diện ban quản trị 46**](#_Toc45002745)

[**4.3 Giao diện đồ thị đánh giá độ chính xác của thuật toán: 52**](#_Toc45002746)

[**CHƯƠNG 5. KẾT LUẬN VÀ KIẾN NGHỊ 55**](#_Toc45002747)

[**5.1 Kết quả đạt được 55**](#_Toc45002748)

[**5.2 Hạn chế của đề tài 55**](#_Toc45002749)

[**5.3 Hướng phát triển của đề tài 56**](#_Toc45002750)

[**5.4 Đề nghị ý kiến 56**](#_Toc45002751)

**Danh mục hình vẽ, đồ thị**

**Danh mục bảng biểu**

**Danh mục các ký hiệu, từ viết tắt**

**3 cái 3 trang.**

# LỜI MỞ ĐẦU

Ngành công nghệ thông tin là một ngành khoa học đang trên đà phát triển mạnh và ứng dụng rộng rãi trên nhiều lĩnh vực. Cùng với xu hướng phát triển của các phương tiện truyền thông như Báo, Radio... thì việc sử dụng Internet ngày càng phổ biến. Truy cập Internet, chúng ta có được một kho thông tin khổng lồ phục vụ mọi nhu cầu, mục đích của chúng ta chỉ bằng một cái nhấp chuột.

Nhận thức được nhu cầu tìm hiểu thông tin, giải trí của xã hội, là sự ra đời của hàng loạt website cho các mục đích thương mại, giải trí, tin tức...Để đáp ứng với việc cập nhật thông tin hàng ngày, tình hình xã hội, chính trị, thời sự, và sức khỏe... thì website tin tức ra đời là một nhu cầu tất yếu.

Do đó mà ở đề tài này tôi quyết định xây dựng một trang web tin tức sử dụng framework Laravel của PHP và cơ sở dữ liệu MySQL.

Ngoài ra, trong ngành khoa học máy tính, trí tuệ nhân tạo hay AI hiện đang phát triển với tốc độ “hàm mũ”, có nhiều đóng góp quan trọng vào sản xuất, kinh doanh, dịch vụ và đời sống con người.

Nhận thấy rằng trí tuệ nhân tạo là một lĩnh vực đang trên đà phát triển rất mạnh mẽ nên tôi quyết định ứng dụng thêm vào đề tài của mình các thuật toán phân loại để nâng cao độ phong phú.

**CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI**

**1.1 Lý do chọn đề tài**

“Tin tức tích lũy trong mỗi chúng ta được gọi là kiến thức, tài sản quý giá nhất của con người”.

Với sự bùng nổ thông tin trên Internet, vai trò của các trang thông tin điện tử trực tuyến càng trở nên quan trọng. Khác với báo chí truyền thông có giới hạn thời gian cập nhật tin tức, các tờ báo trực tuyến đã cung cấp được sự tiện lợi trong việc cập nhật và phát hành thông tin. Về phía người dùng, họ có thể xem thông tin mọi lúc mọi nơi. Về phía những người cung cấp thông tin, các nhà báo, họ có thể dễ dàng cập nhật những tin tức mới nhất, thời sự nhất. Do đó việc sử dụng các trang thông tin trực tuyến luôn là điều cần thiết hiện nay nhằm đáp ứng nhu cầu cập nhật thông tin của mỗi người.

Tin tức là những việc đã xãy ra dù tốt dù xấu, để giúp con người biết những chuyện xung quanh và trên Thế giới. Ngày nay nhờ thông tin truyền thông nhanh, cho nên bất cứ chuyện gì vừa xãy ra ở đâu trên thế giới thì ta đều có thể biết ngay, nhờ đó mà có thể học được nhiều cái hay cũng như tránh được những chuyện xấu xảy ra, như các trận sóng Thần, bão táp, núi lửa sắp đến, các chất độc hại trong thức ăn.... giúp con người biết trước mà tránh khỏi các nguy hiểm sắp đến.

Tin tức vô cùng quan trọng nó cho người ta tri thức và là cơ sở để người ta tiến hành mọi việc lớn nhỏ. Khi có Internet, tin tức càng quan trọng vì tốc độ lan truyền nhanh ảnh hưởng ngay tức thì trên diện rộng.

Vậy nên tôi muốn xây dựng một trang web tin tức để có thể đưa đến cho tất cả mọi người những thông tin nhanh và chính xác nhất. Nhưng tin tức là thứ luôn phải được cập nhật hằng ngày, hằng giờ nên tôi quyết định sẽ lấy dữ liệu tin tức từ các trang web tin tức nổi tiếng hiện nay như Báo mới, VnExpress, Dân trí, … để có thể tự động cập nhật tin tức đến cho mọi người.

Nhận thấy trí tuệ nhân tạo (Artificial Intelligence) và học máy (machine learning) ngày càng được ứng dụng rộng rãi trên nhiều lĩnh vực hiện nay nên tôi quyết định tìm hiểu thêm thuật toán phân loại để ứng dụng cho trang web nhằm mục đích phân loại các dữ liệu tin tức vào các thể loại.

Khi xây dựng trang web tin tức, có thể sử dụng các phầm mềm nguồn mở, việc cài đặt sẽ nhanh hơn rất nhiều, tuy nhiên khuyết điểm của các phần mềm nguồn mở là khó tùy biến theo nhu cầu của mình. Cách khác để xây dựng website là tự lập trình để xây dựng hệ thống, có thể bắt đầu từ đầu và tất nhiên sẽ rất lâu, hay lựa chọn các thư viện phù hợp như Laravel, Zend, CodeIgniter… giúp xây dựng hệ thống nhanh hơn. Sau khi tìm hiểu qua các thư viện, tôi quyết định chọn Laravel Framework vì những tiện lợi của Laravel mang đến và cảm thấy nó phù hợp với đồ án này.

Do đó, tôi chọn thực hiện đề tài *“Ứng dụng thuật toán phân loại văn bản trong xây dựng trang web tin tức sử dụng Laravel và NodeJS”*

Hệ thống được xây dựng với định hướng giúp mọi người có thể cập nhật thông tin một cách nhanh chóng. Từ đó nâng cao nhận thức của bản thân về thế giới xung quanh. Đồng thời góp phần giúp cho chúng ta chuẩn bị mọi thứ tốt hơn trong cuộc sống.

**1.2 Đối tượng và phạm vi nghiên cứu**

* + - Hệ thống trang web tin tức.
    - Các ngôn ngữ, nền tảng sử dụng:
      * HyperText Markup Language (HTML)
      * Cascading Style Sheets (CSS)
      * JavaScript (JS)
      * Hypertext Preprocessor (PHP)
    - Các framework: Laravel cho trang web
    - Nền tảng: NodeJS ứng dụng thuật toán để xử lý dữ liệu
    - Mô hình Model – View – Controller (MVC) của Laravel.
    - Cơ sở dữ liệu: MySQL.
    - Thuật toán phân lớp Naive Bayes

**1.3 Mục đích nghiên cứu**

Xây dựng trang web tin tức lấy nguồn tin từ các trang web tin tức tin cậy hiện nay như: Kenh14, tuoitre.vn, VnExpress…, các tin tức khi đưa vào hệ thống xây dựng của website sẽ được tự động phân loại vào các nhóm tin dựa trên thuật toán phân loại.

**1.4 Nội dung thực hiện**

**1.4.1 Về trang web:**

* + - Lấy dữ liệu tin tức từ các trang web tin tức nổi tiếng hiện nay như Kenh14, dantri.com.vn , tuoitre.vn…
    - Quản lý các thông tin về loại tin, tin tức, người dùng, ban quản trị.
    - Hiển thị các tin tức có tính thời sự, nội dung chính xác, nhanh chóng và hấp dẫn cho mọi thể loại lứa tuổi.
    - Cung cấp cho người dùng những cập nhật mới nhất về tin tức hiện nay.
    - Sắp xếp, bố cục một cách hợp lý, tạo điều kiện cho người đọc dễ định hướng và tìm ra thể loại tin tức mà mình muốn đọc trong trang web.
    - Hình thức đẹp phù hợp với chủ đề, nội dung, thân thiện với bạn đọc.

**1.4.2 Về thuật toán:**

* + - Hiểu được thuật toán phân loại Naive Bayes Classifier.
    - Ứng dụng nó vào hệ thống nhằm mục đích phân loại tin tức.
    - Đưa ra được các thống kê, báo cáo, độ chính xác về việc phân loại tin tức dựa vào thuật toán trên.

**1.4.3 Về dữ liệu :**

* Tiến hành thu thập dữ liệu.
* Tiền xử lý dữ liệu cho dữ liệu mẫu và dữ liệu thử nghiệm.
* Tiến hành dự đoán và phân loại dữ liệu.
* Lưu dữ liệu vào cơ sở dữ liệu.

**1.5 Môi trường cài đặt và IDE thực hiện đề tài**

Tôi sử dụng IDE (Integrated Development Environment) là Visual Studio Code được cài đặt trên môi trường Windows, cụ thể là Windows 10 Pro (64-bit) để thực hiện đề tài này.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## **2.1 Khái quát về HTML**

HTML là chữ viết tắt của cụm từ HyperText Markup Language, dịch sang tiếng việt có nghĩa là ngôn ngữ đánh dấu siêu văn bản, nó được dùng để tạo ra trang web (tất cả những nội dung mà chúng ta nhìn thấy trên trang web, ví dụ như: văn bản, hình ảnh, liên kết, . . . . chính là được tạo ra dựa trên các thẻ HTML)

## **2.2 Khái quát về CSS**

CSS là ngôn ngữ tạo phong cách cho trang web – Cascading Style Sheet language. Nó dùng để tạo phong cách và định kiểu cho những yếu tố được viết dưới dạng ngôn ngữ đánh dấu, như là HTML. Nó có thể điều khiển định dạng của nhiều trang web cùng lúc để tiết kiệm công sức cho người viết web. Nó phân biệt cách hiển thị của trang web với nội dung chính của trang bằng cách điều khiển bố cục, màu sắc, và font chữ.

## **2.3 Khái quát về JavaScript**

JS (viết tắt của Javascript) là một nền tảng (cross-platform), ngôn ngữ kịch bản hướng đối tượng (object-oriented). Nó là một ngôn ngữ nhỏ và nhẹ. Chạy trong môi trường máy chủ lưu trữ (ví dụ: trình duyệt web), JavaScript có thể được kết nối với các đối tượng của môi trường để cung cấp kiểm soát chương trình đối với chúng.

JavaScript cho phép bạn thực hiện những điều phức tạp trên các trang web như bản đồ tương tác...

## **2.4 Khái quát về MySQL**

MySQL là một hệ thống quản lý cơ sở dữ liệu quan hệ mã nguồn mở (RDBMS) dựa trên ngôn ngữ truy vấn có cấu trúc ( SQL) được phát triển, phân phối và hỗ trợ bởi tập đoàn Oracle. MySQL chạy trên hầu hết tất cả các nền tảng, bao gồm cả Linux , UNIX và Windows. MySQL thường được kết hợp với các ứng dụng web.

## **2.5 Khái quát về PHP**

PHP: Hypertext Preprocessor, thường được viết tắt thành PHP là một ngôn ngữ lập trình kịch bản hay một loại mã lệnh chủ yếu được dùng để phát triển các ứng dụng viết cho máy chủ, mã nguồn mở, dùng cho mục đích tổng quát. Nó rất thích hợp với web và có thể dễ dàng nhúng vào trang HTML. Do được tối ưu hóa cho các ứng dụng web, tốc độ nhanh, nhỏ gọn, cú pháp giống C và Java, dễ học và thời gian xây dựng sản phẩm tương đối ngắn hơn so với các ngôn ngữ khác nên PHP đã nhanh chóng trở thành một ngôn ngữ lập trình web phổ biến nhất thế giới.

## **2.6 Framework Laravel**

### **2.6.1 Khái niệm**

Laravel là một PHP framework [1] mã nguồn mở và miễn phí, được phát triển bởi Taylor Otwell và nhắm vào mục tiêu hỗ trợ phát triển các ứng dụng web theo kiếm trúc model-view-controller (MVC). Những tính năng nổi bật của Laravel bao gồm cú pháp dễ hiểu – rõ ràng , một hệ thống đóng gói modular và quản lý gói phụ thuộc, nhiều cách khác nhau để truy cập vào các cơ sở dữ liệu quan hệ, nhiều tiện ích khác nhau hỗ trợ việc triển khai vào bảo trì ứng dụng.

Vào khoảng Tháng 3 năm 2015, các lập trình viên đã có một cuộc bình chọn PHP framework phổ biến nhất, Laravel đã giành vị trí quán quân cho PHP framework phổ biến nhất năm 2015, theo sau lần lượt là Symfony2, Nette, CodeIgniter, Yii2 vào một số khác. Trước đó, Tháng 8 2014, Laravel đã trở thành project PHP phổ biến nhất và được theo dõi nhiều nhất trên Github.

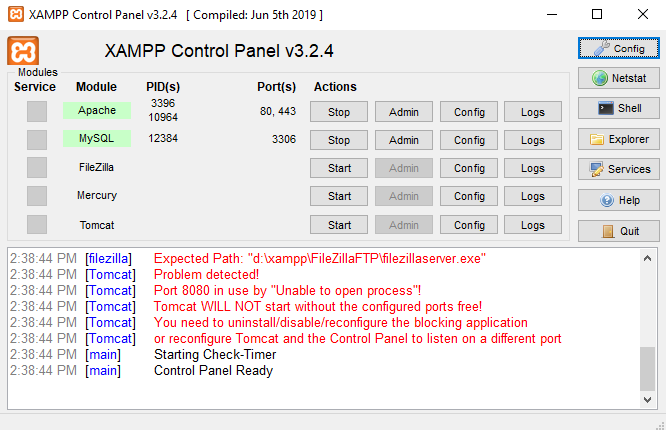
### **2.6.2 Cấu hình yêu cầu khi cài đặt Laravel**

* PHP >= 5.6.4
* OpenSSL PHP Extension
* PDO PHP Extension
* Mbstring PHP Extension
* Tokenizer PHP Extension
* XML PHP Extension

### **2.6.3 Cài đặt Laravel**

* Cài đặt Xampp:

Truy cập đường link <https://www.apachefriends.org/download.html>, tải về vài tến hành cài đặt, sau khi cài đặt thành công thì Xampp sẽ có giao diện như vậy

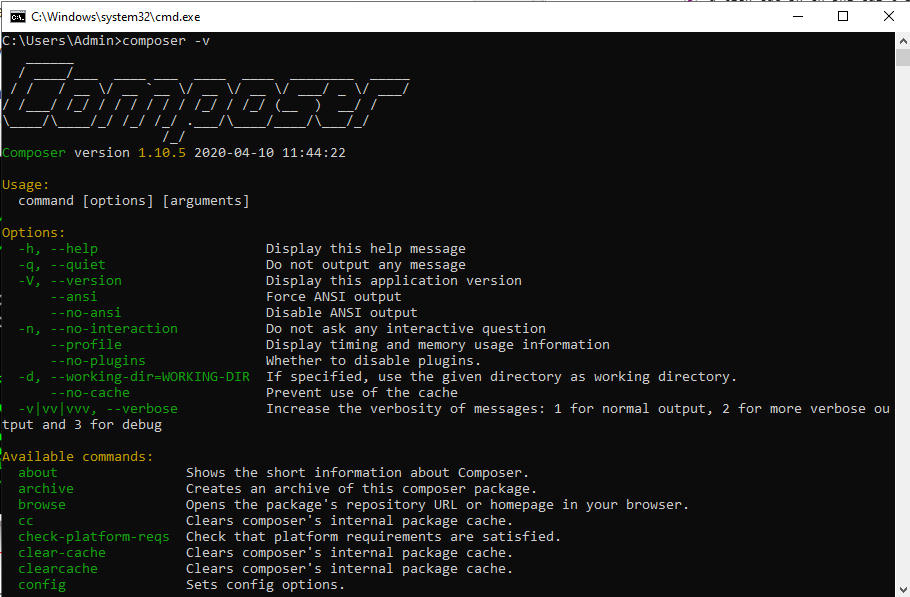


Hình 2.1 Giao diện Xampp Control Panel

* Cài đặt Composer:

Đầu tiên vào link <https://getcomposer.org/Composer-Setup.exe> và tải về.

Sau quá trình cài đặt thì ta vào command line của Windows (windows + R open cmd) gõ “composer -v”. Màn hình hiện ra như này tức là đã cài đặt composer thành công.



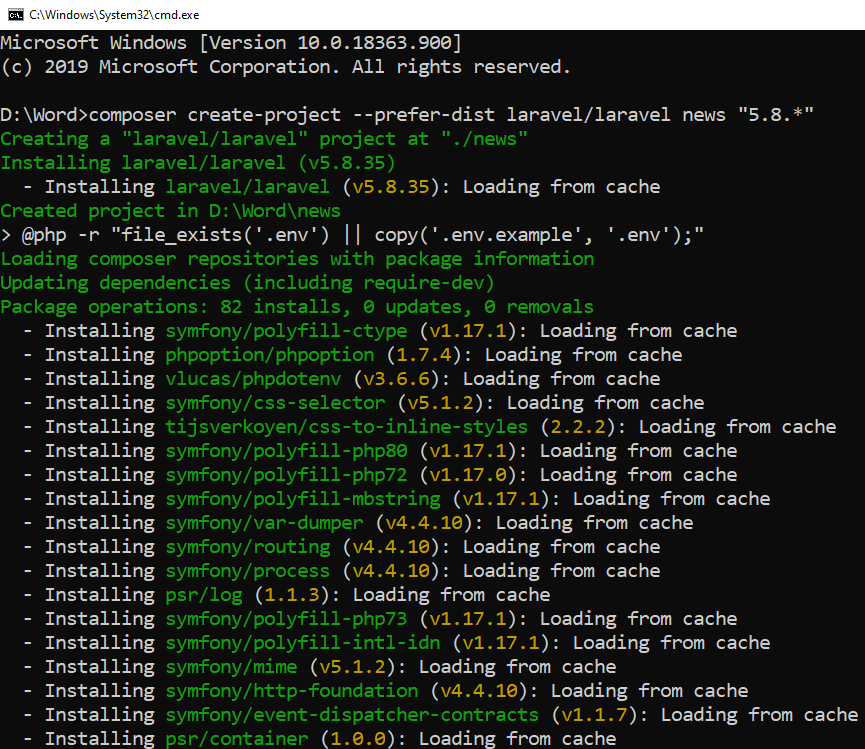
Hình 2.2 Composer trên giao diện dòng lệnh CLI

* Cuối cùng là cài đặt một project Laravel chỉ bằng một câu lệnh:

*composer create-project --prefer-dist laravel/laravel news "5.8.\*"*

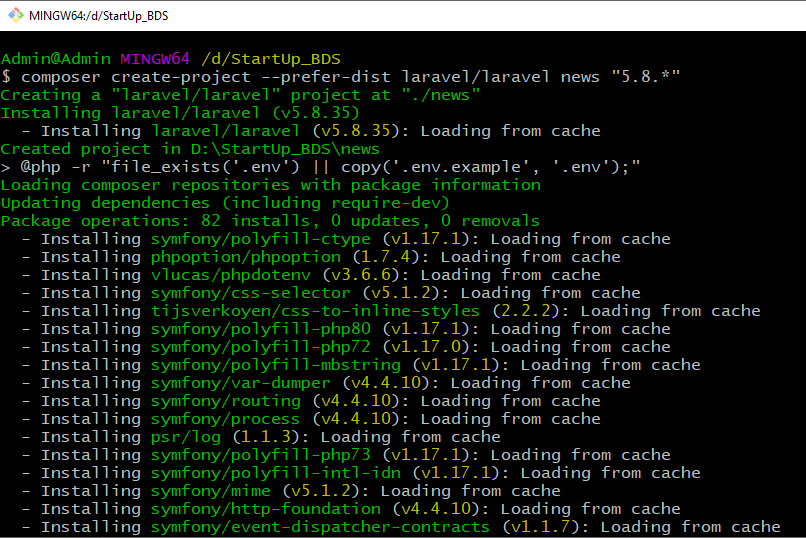
Trong đó news là tên của thư mục Laravel project, còn “5.8.\*” là phiên bản của project.

Chúng ta có thể chạy câu lệnh trên bằng cửa sổ cmd



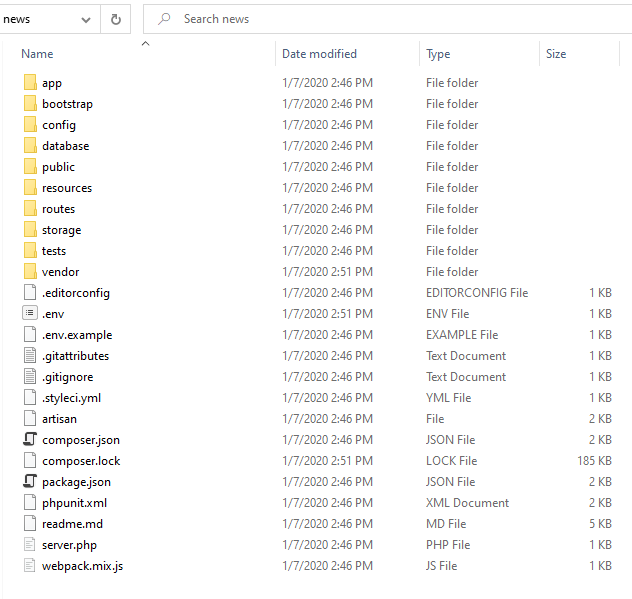
Hình 2.3 Màn hình cài đặt Laravel trên cửa sổ cmd

hoặc Git Bash



Hình 2.4 Màn hình cài đặt Laravel trên Git Bash

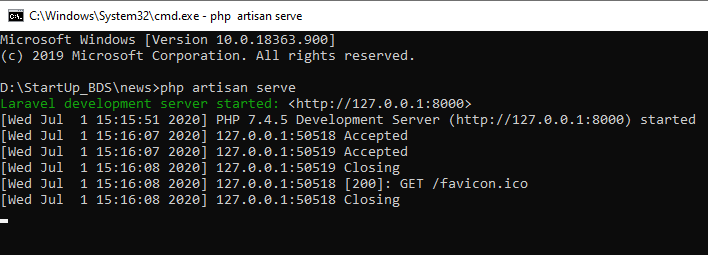
Đợi một lát và sau đây là thành quả. Một thư mục chứa hàng loạt thư mục con và file.



Hình 2.5 Thư mục và tập tin của một project Laravel

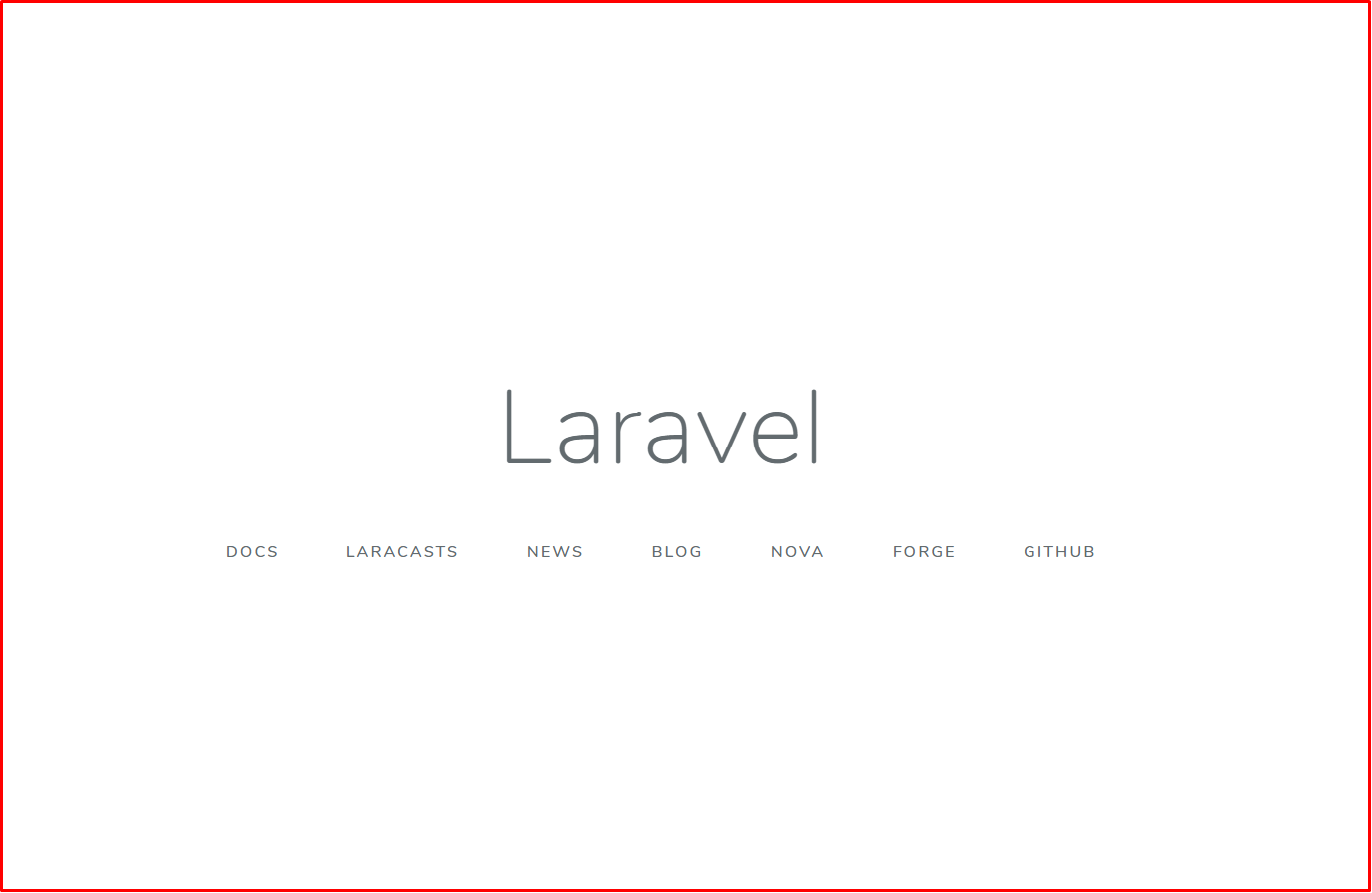
* Sau khi cài đặt hoàn tất, đi đến thư mục Laravel project và chạy câu lệnh sau để khởi động Server:

*php artisan serve*



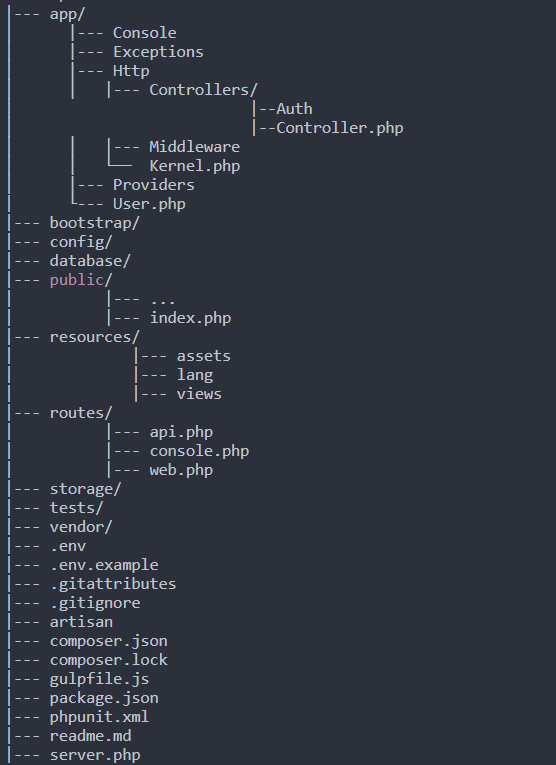
Hình 2.6 Giao diện chạy câu lệnh khởi tạo server Laravel

Sau khi chạy dòng lệnh, mở trình duyệt và truy cập địa chỉ <http://localhost:8000> hoặc <http://127.0.0.1:8000>, chúng ta sẽ thu được kết quả như hình bên dưới:



Hình 2.7 Giao diện mặc định khi lần đầu chạy project Laravel

### **2.6.4 Cấu trúc thư mục của một project Laravel:**



Hình 2.8 Cấu trúc cây thư mục Laravel

Chú thích [2]:

* app: Thư mục này rất quan trọng và bạn thường xuyên phải thao tác trong thư mục này. Nó chứa các Models, Controller, Providers ...
* Console: Chứa các tập tin định nghĩa các câu lệnh trên artisan.
* Excerption: Chứa các tập tin quản lý, điều hướng lỗi.
* Http:
* Controllers: Chứa các controllers của project.
* Middleware: Chứa các tập tin lọc và ngăn chặn các requests.
* Kernel.php: Cấu hình, định nghĩa Middleware.
* Providers: Chứ các providers thực hiện việc binding vào service container(ở phần nâng cao mình sẽ nói sau).
* User.php: Là model User mà Laravel tự địn sẵn cho chúng ta.
* bootstrap: Chứa các file khởi tạo và một số file cấu hình nạp tự động.
* config: Chứa các file cấu hình của Laravel Framework: cấu hình database, app, mail, session ...
* database: Chứa các thư mục tập tin về database.
* migrations: Chứa các tập tin định nghĩa,khởi tạo và sửa bảng.
* seeds: Chứa các tập tin định nghĩa dữ liệu insert(thêm) vào trong database.
* factories: Chứa các tập tin định nghĩa các cột bảng dữ liệu để tạo ra các dữ liệu ảo.
* public: Các file tĩnh như file HTML, CSS, Javascript hay ảnh sẽ được đưa vào thư mục này.
* index.php: Đây là tệp tin root của Laraver
* resources: Chứa trang views (V của mô hình MVC), chứa các file khác như (SASS, LESS, CoffeeScript) và các file hỗ trợ cho đa ngôn ngữ.
* routes: Chứa các tập tin định nghĩa các router, xử lý điều hướng router bao gồm : web, api và console.
* storage: Chứa các file đã biên dịch từ các file view xài Blade template, chứa file sessions, caches và các file được sinh ra tự động của framework.
* tests: Chứa các file test của Framework.
* vendor: Chứa các thư viện mà Composer cài đặt cho Framework.
* .env: Là tập tin cấu hình chính của laravel như key app, database.
* .env.example: Tệp tin cấu hình mẫu của laraver.
* composer.json: Chứa danh sách các dependencies (thư viện, framework) sử dụng trong app.
* package.js: Tập tin cấu hình của nodejs (chứa các package cần dùng cho projects).
* gulpfile.js: Là tập tin gulp builder.
* phpunit.xml: Là tập tin xml của phpunit dùng để testing project.
* server.php: Là tập tin để artisan trỏ đến tạo server khi gõ lệnh php artisan serve .
* artisan: Được dùng để thực thi các lệnh CLI (command line interface) để hỗ trợ phát triển ứng dụng..

## **2.7 NodeJS**

### **2.7.1 Khái niệm**

Nodejs [3] là một **nền tảng** (Platform) phát triển độc lập được xây dựng ở trên Javascript Runtime của Chrome mà chúng ta có thể xây dựng được các ứng dụng mạng một cách nhanh chóng và dễ dàng mở rộng. Nodejs được xây dựng và phát triển từ năm 2009, bảo trợ bởi công ty Joyent, trụ sở tại California, Hoa Kỳ. Phần Core bên dưới của Nodejs được viết hầu hết bằng C++ nên cho tốc độ xử lý và hiệu năng khá cao.

Nodejs tạo ra được các ứng dụng có tốc độ xử lý nhanh, realtime thời gian thực. Nodejs áp dụng cho các sản phẩm có lượng truy cập lớn, cần mở rộng nhanh, cần đổi mới công nghệ, hoặc tạo ra các dự án Startup nhanh nhất có thể.

### **2.7.2 Những ứng dụng nên viết bằng Nodejs**

Rõ ràng, không phải cứ hot và mới là Nodejs làm gì cũng tốt, ví dụ như một ứng dụng cần tính ổn định cao, logic phức tạp thì các ngôn ngữ PHP hay Ruby… vẫn là sự lựa chọn tốt hơn. Còn dưới đây là những ứng dụng có thể và nên viết bằng Nodejs:

* **Websocket server:** Các máy chủ web socket như là Online Chat, Game Server…
* **Fast File Upload Client:** là các chương trình upload file tốc độ cao.
* **Ad Server:** Các máy chủ quảng cáo.
* **Cloud Services:** Các dịch vụ đám mây.
* **RESTful API:** đây là những ứng dụng mà được sử dụng cho các ứng dụng khác thông qua API.
* **Any Real-time Data Application:** bất kỳ một ứng dụng nào có yêu cầu về tốc độ thời gian thực. Micro Services: Ý tưởng của micro services là chia nhỏ một ứng dụng lớn thành các dịch vụ nhỏ và kết nối chúng lại với nhau. Nodejs có thể làm tốt điều này.

### **2.7.3 Những kết luận sai lầm về Nodejs**

* Thứ nhất, Nodejs là một nền tảng (platform), không phải Web Framework, cũng không phải ngôn ngữ lập trình.
* Thứ hai, Nodejs không hỗ trợ đa luồng, nó là một máy chủ đơn luồng.
* Và một điều nữa, Nodejs không dành cho người mới tinh mà chưa biết gì về lập trình, vì như đã nói ở trên, Nodejs không phải là ngôn ngữ lập trình, để học được Nodejs thì bạn cần phải biết về Javascript, kỹ thuật lập trình, một số giao thức…

### **2.7.4 Lý do nên sử dụng Nodejs**

* Các ứng dụng Nodejs được viết bằng javascript, ngôn ngữ này là một ngôn ngữ khá thông dụng. Theo tác giả của ngôn ngữ Javascript, Ryan Dahl: “Javascript có những đặc tính mà làm cho nó rất khác biệt so với các ngôn ngữ lập trình động còn lại, cụ thể là nó không có khái niệm về đa luồng, tất cả là đơn luồng và hướng sự kiện.”
* Nodejs chạy đa nền tảng phía Server, sử dụng kiến trúc hướng sự kiện Event-driven, cơ chế non-blocking I/O làm cho nó nhẹ và hiệu quả.
* Có thể chạy ứng dụng Nodejs ở bất kỳ đâu trên máy Mac – Window – Linux, hơn nữa cộng đồng Nodejs rất lớn và hoàn toàn miễn phí. Các bạn có thể thấy cộng đồng Nodejs lớn như thế nào tại đây, các package đều hoàn toàn free: <https://www.npmjs.com/>
* Các ứng dụng NodeJS đáp ứng tốt thời gian thực và chạy đa nền tảng, đa thiết bị.

### **2.7.5 Node Package Manager (NPM)**

NPM viết tắt của Node Package Manager là một công cụ tạo và quản lý các thư viện lập trình Javascript cho Node.js. Trong cộng đồng Javascript, các lập trình viên chia sẻ hàng trăm nghìn các thư viện với các đoạn code đã thực hiện sẵn một chức năng nào đó. Nó giúp cho các dự án mới tránh phải viết lại các thành phần cơ bản, các thư viện lập trình hay thậm chí cả các framework.

Cộng dụng của npm làm cho công việc sẽ đơn giản đi rất nhiều, chúng giúp thực hiện việc quản lý đơn giản hơn rất nhiều. Các thư viện đều có sẵn trên npm, bạn chạy một dòng lệnh để tải về và dễ dàng include chúng hơn.

### **2.7.6 Cài đặt NodeJS**

Để download NodeJS, chỉ cần truy cập vào link: <https://nodejs.org/en/download/>, lựa chọn phiên bản phù hợp sau đó tải xuống và tiến hành cài đặt.

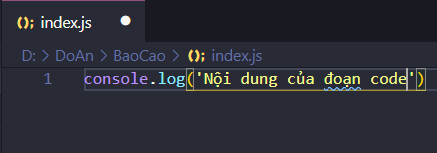
Theo mặc định, phần mềm NPM cũng được cài đặt vào hệ thống. Đây là một phần mềm quản lý các thư viện Javascript.

Sau khi hoàn tất quá trình cài đặt, cần kiểm tra lại kết quả và cấu hình NodeJS. Mở cửa sổ cmd và thực thi hai câu lệnh “node -v” và “npm -v” để xem phiên bản của NodeJS và NPM. Nếu kết quả hiển thị không phải là phiên bản của node và npm thì tức là quá trình cài đặt chưa thành công. Hiện tại tôi đang sử dụng “node v13.12.0” và “npm 6.14.4”

### **2.7.7 Sử dụng NodeJS cơ bản**

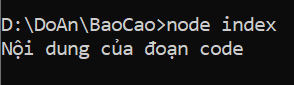
Để chạy NodeJS rất đơn giản

* Đầu tiên tạo một file javascript bất kì (ví dụ là index.js)
* Tiếp đến gõ một đoạn code đơn giản để có thể hiện thị nội dung.



Hình 2.9 Đoạn code đơn giản trong Visual Code

* Cuối cùng là mở cửa số cmd của thư mục chứa file index.js gõ: “node index”, và đây là kết quả:



Hình 2.10 Kết quả khi thực thi đoạn code

## **2.8 Các module xử lý dữ liệu và ứng dụng thuật toán Naive Bayes vào hệ thống**

* **Mysql** [4]**:** là một module trong NodeJS giúp kết nối đến cơ dữ liệu. Sử dụng module này nhằm mục đích lưu trữ dữ liệu được lấy về từ các trang web khác vào cơ sở dữ liệu.
* Cách cài đặt module này về dự án rất đơn giản chỉ mới một câu lệnh:

“*npm install mysql*”

* Sau khi được tải về module này sẽ được lưu trữ trong thư mục node\_modules.
* Và khi muốn sử dụng thì chỉ cần lấy nó ra bằng cách như sau:

“*const mysql = require(‘mysql’)*”

* **Puppeteer** [5][6]**:** Puppeteer là một ‘Node library’ do Google phát triển cung cấp các API điều khiển Chrome hoặc Chromium thông qua DevTools Protocol. Puppeteer mặc định chạy ở chế độ headless tuy nhiên ta cũng có thể cài đặt để chạy non-headless. Hầu hết những việc có thể thực hiện thủ công trên trình duyệt đều có thể được thực hiện bằng Puppeteer. Mục đích sử dụng puppeteer là để lấy được dữ liệu từ các trang web tin tức khác.
* **Remove-punctuation** [7]**:** module này giúp xoá tất cả các dấu câu. Đây là một phần trong bước tiền xử lý dữ liệu trước khi phân loại.

**Vietnamese-stopwords** [8]**:** module này giúp xoá tất cả những từ thuộc stopword của tiếng Việt (những từ vô nghĩa trong câu). Đây cũng là một phần trong bước tiền xử lý dữ liệu trước khi phân loại.

* **Vntk** [9]**:** module này được ví như bộ công cụ xử lý ngôn ngữ tự nhiên dành cho tiếng Việt. Nhằm mục đích hỗ trợ, giúp tôi có thể tạo ra các sản phẩm có liên quan đến xử lý ngôn ngữ tự nhiên bằng Nodejs. Đặc biệt là tiếng Việt.
* **WhichX** [10]**:** là một thư viện nhỏ với API (Application Programming Interface) đơn giản được dùng để phân loại văn bản vào các danh mục có sẵn dựa trên thuật toán phân loại Naive Bayes Classifier, được viết bằng ngôn ngữ lập trình JavaScript. Đây là thư viện quan trọng nhất trong đề tài, vì nó được dùng trong việc ứng dụng trực tiếp thuật toán phân lớp Naive Bayes Classifier nhằm mục đích phân loại các tin tức vào các thể loại cho sẵn.
* Cách cài đặt thư viện này rất đơn giản chỉ mới một câu lệnh:

“*npm install whichx*”

* Sau khi được tải về thư viện này sẽ được lưu trữ trong thư mục node\_modules, dưới dạng là một thư mục có tên là *whichx.* Ở thư mục này có nhiều tập tin và thư mục con để cấu tạo nên một thư viện. Trong đó có tập tin *whichx.js* là tập tin quan trọng nhất vì nó là tập tin chính trong thư mục *whichx* giúp cho thư viện này thực hiện các chức năng phân loại. Cụ thể sẽ phân tích các chức năng (function) chính trong tập tin này:
* Function này có chức năng thêm nhãn (labels) hoặc danh sách nhãn vào phân loại

*this.addLabels = function(labels)*

Ví dụ:

var labels = ['pháp\_luật', 'kinh\_doanh', 'công\_nghệ', 'xe', 'giáo\_dục', 'sức\_khoẻ'];

category.addLabels(labels);

* Function này có chức năng thêm dữ liệu từ mô tả (description) vào các nhãn đã được chỉ định.

*this.addData = function(label, description)*

Ví dụ:

category.addData('pháp\_luật', arrayTrains[0]);

category.addData('kinh\_doanh', arrayTrains[1]);

category.addData('công\_nghệ', arrayTrains[2]);

category.addData('xe', arrayTrains[3]);

category.addData('giáo\_dục', arrayTrains[4]);

category.addData('sức\_khoẻ', arrayTrains[5]);

với biến arrayTrains là mảng chứa dữ liệu của từng thể loại

* Function này sẽ lấy mô tả và tìm nhãn có khả năng được phân loại vào đó nhất.

*this.classify = function(description)*

Ví dụ:

var News = category.classify(total\_array\_testing[array][i];

với biến total\_array\_testing là dữ liệu của tập dự đoán đưa vào để phân lớp

Ngoài ra trong tập tin *whichx.js* này còn nhiều function với chức năng khác nhau như loại bỏ stopword, chuẩn hoá từ, … nhưng tôi đã xử lý các chức năng này ở phần khác nên ở đây sẽ không nhắc đến nữa.S

* Và khi muốn sử dụng thì chỉ cần gọi nó ra bằng cách như sau:

“*const WhichX = require(‘whichx’)*”

## **2.9 Định lý Bayes**

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

* Theo định lí Bayes [11, 12], xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:
* Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là P(A) và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm, nó là “tiên nghiệm” theo nghĩa rằng nó không quan tâm đến bất kỳ thông tin nào về B.
* Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là P(B) và đọc là “xác suất của B”. Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.
* Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là P(B|A) và đọc là “xác suất của B nếu có A”. Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra B khi biết A và xác suất xảy ra A khi biết B.
* Tóm lại định lý Bayes sẽ giúp ta tính ra xác suất xảy ra của một giả thuyết bằng cách thu thập các bằng chứng nhất quán hoặc không nhất quán với một giả thuyết nào đó. Khi các bằng chứng tích lũy, mức độ tin tưởng vào một giả thuyết thay đổi. Khi có đủ bằng chứng, mức độ tin tưởng này thường trở nên rất cao hoặc rất thấp, tức là xác xuất sảy ra giả thuyết sẽ thay đổi thì các bằng chứng liên quan đến nó thay đổi.
* Công thức Bayes:

Với P(B) > 0:

Suy ra:

Công thức Bayes:

Trong đó:

* + P(A|B) là xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.
  + P(B|A) là xác suất xảy ra B khi biết A xảy ra
  + P(A) là xác suất sảy ra của riêng A mà không quan tâm đến B.
  + P(B) là xác suất xảy ra của riêng B mà không quan tâm đến A.

Ở trên ta có thể thấy xác suất xảy ra của giả thuyết A phụ thuộc và xác suất của giả thuyết B, nhưng trong thực tế xác suất A có thể phụ thuộc vào xác suất của nhiều các giác thuyết khác có thể là B1, B2, B3 … Bn. Vậy định luật Bayes có thể được mở rộng bằng công thức tổng quát sau:

Với P(A) > 0 và {B1, B2, …, Bn} là một hệ đầy đủ các biến cố:

* Tổng xác suất của hệ bằng 1:

* Từng đôi một xung khắc:

Khi đó ta có:

**Ví dụ:** ta có một thống kê như sau:

Trong Quốc hội Mỹ có 200 thượng nghị sĩ trong đó có:

+ 120 thượng nghị sĩ thuộc đảng Dân Chủ.

+ 80 thượng nghĩ sĩ thuộc đảng Cộng Hòa.

+ Số lượng nữ giới trong Quốc hội là 60 người

+ Còn lại 140 người là nam giới.

+ Và số lượng nữ giới trong đảng Dân Chủ là 30 người.

Vậy nếu tôi chọn ngẫu nhiên một người trong Quốc hội thì tỷ lệ chọn đúng thượng nghị sĩ là nữ giới và thuộc đảng Dân Chủ là bao nhiêu phần trăm?

Áp dụng công thức Bayes ta có thể tính toán được bằng công thức sau:

• P(Female|Democrat): Chính là tỷ lệ nữ giới thuộc đảng Dân chủ trong cả Quốc hội cần tính toán.

• P(Demorate|Female): Chính là tỷ lệ nữ giới trong đảng Dân chủ.

• P(Female): Chính là tỷ lệ nữ giới trong Quốc hội.

• P(Democrat): Chính là tổng số thượng nghị sĩ trong Quốc hội.

Ở đây với dữ liệu cho bên trên ta có thể tính toán được:

P(Democrat|Female) = Số nữ giới trong đảng Dân chủ / Tổng số thượng nghị sĩ của đảng Dân chủ.

P(Democrat|Female) = 30/ 120 = 0.25

P(Female) = Số nữ giới trong Quốc hội / Tổng số thượng nghị sĩ trong Quốc hội

P(Female) = 60/200 = 0.3

P(Democrat) = Tổng số thượng nghĩ sĩ trong Quốc hội.

P(Democrat) = 1

Vậy ta có thể tín ra P(Female|Democrat) theo công thức Bayes như sau:

P(Female|Democrat) = (0.25 \* 0.3) / 1 = 0.075

Có nghĩa là nếu tôi chọn chọn ngẫu nhiên một người trong số thượng nghị sĩ thì tỷ lệ thượng nghị sĩ là Nữ giới và thuộc đảng Dân Chủ sẽ là “7,5%”.

## **2.10 Thuật toán phân lớp Naive Bayes Classification**

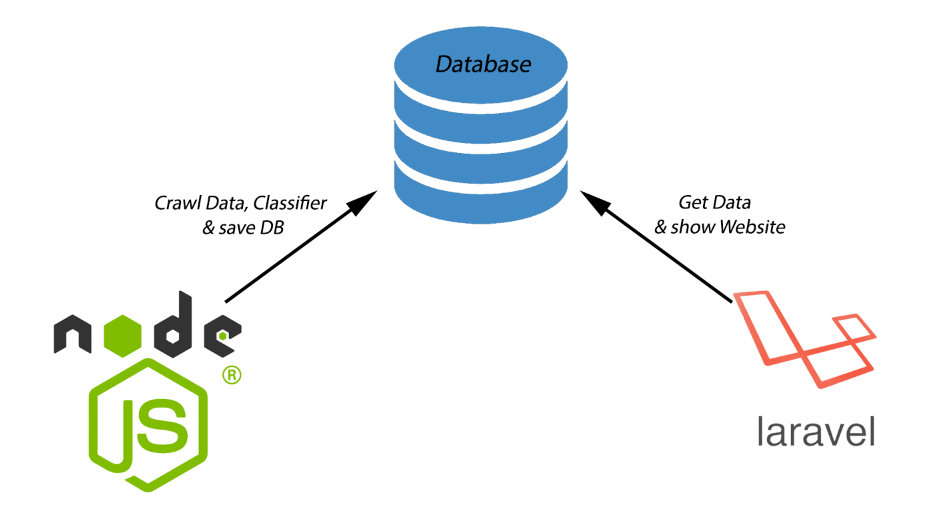
Naive Bayes Classification (NBC) (2 cái link như bayees luôn) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất  để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các tập dữ liệu đã được xử lý trước đó. Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có giám sát), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.

Ví dụ như ta có thể ứng dụng vào việc thiết kế một ứng dụng nghe nhạc có thể phán đoán được sở thích của nghe nhạc của người dùng dựa trên các hành vi như nhấn nút “thích” bài hát, “nghe đi nghe” lại nhiều lần các bài hát,  “bỏ qua” các bài hát không thích …. Dựa trên tập dữ liệu đó ta có thể áp dụng NBC để tính toán ra các phong cách nhạc mà người dùng thích nhất, từ đó chúng ta có thể đưa ra các “gợi ý” nghe nhạc gần đúng nhất cho người dùng từ việc học hỏi từ những thói quen đó.

## **2.11 Laravel kết hợp NodeJS**

Trong đề tài này tôi sử dụng NodeJS và Laravel để xử lý mọi công việc ở phần Backend. Tuy đều được sử dụng ở Backend nhưng mỗi cái sẽ đảm nhiệm một chức năng khác nhau. Cụ thể:

* NodeJS sẽ thu thập dữ liệu tin tức và phân loại chúng bằng thuật toán phân loại Naive Bayes Classifier. Sau đó lưu tất cả dữ liệu thu thập được vào trong cơ sở dữ liệu.
* Laravel sẽ có nhiệm vụ lấy dữ liệu đã được phân loại và lưu vào cơ sở dữ liệu để hiển thị trên giao diện website.



Hình 2.11 Mối liên hệ giữa NodeJS và Laravel thông qua Database

Tuy mỗi cái đảm nhiệm mỗi nhiệm vụ khác nhau nhưng giữa chúng có sự liên kết rất chặt chẽ và không thể thiếu một trong hai cái được. Ở đây là việc chúng dùng chung một cơ sở dữ liệu. NodeJS muốn đưa dữ liệu lên website thì phải thông qua Laravel và ngược lại Laravel muốn có dữ liệu để đưa lên website thì phải nhờ đến NodeJS. Nhìn một cách khái quát thì NodeJS đảm nhiệm chức năng nạp dữ liệu vào còn Laravel có nhiệm lấy dữ liệu ra.

# CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

## **3.1 Phân tích và thiết kế hệ thống WebSite**

### **3.1.1 Định nghĩa bài toán**

Như mục tiêu của đề tài và giải pháp đặt ra để khắc phục các hạn chế của phương pháp truyền tin cũ qua các tờ báo. Ta có thể thấy được yêu cầu của bài toán là xây dựng được một hệ thống trang web tin tức để có thể đưa đến cho tất cả mọi người những thông tin nhanh và chính xác nhất … đáp ứng những nhu cầu sau:

* + *Đối với nhà quản trị:* hệ thống cung cấp những chức năng cần thiết để tiến hành chỉnh sửa, quản lý sự hoạt động cũng như theo dõi tình hình phát triển của hệ thống.
  + *Đối với người dùng:* hệ thống cho phép xem chi tiết tin tức của loại tin trên trang web. Sau khi đăng ký làm thành viên, người dùng có thể chọn xem tin tức của từng loại tin, họ có thể đánh giá tin tức đó thông qua việc bình luận ở phần dưới của bài đăng. Khi người dùng muốn nhanh chóng tìm kiếm một thông tin mà họ muốn thì họ có thể nhập từ khoá trên thanh tìm kiếm, nhằm tiết kiệm thời gian tìm kiếm thông tin.

### **3.1.2 Sơ đồ phân ra chức năng**

Hệ thống được chia làm 2 phần: front-end (dành cho người dùng) và back-end (dành cho người quản trị hệ thống)

#### **3.1.2.1 Front-end**

Từ mô tả, định nghĩa bài toán xây dựng trang web tin tức, ta có thể xác định một số chức năng sẽ có dành cho người dùng:

* **Các chức năng:**

Bảng 3.1. Bảng chức năng của Front-end

|  |  |  |
| --- | --- | --- |
| **STT** | **Chức năng** | **Mô tả** |
| 1 | Xem tin | Người xem có thể xem thông tin chi tiết về tin tức đó bao gồm: tiêu đề, mô tả, hình và phần nội dung. |
| 2 | Bình luận | Chức năng bình luận cho phép người dùng có thể đánh giá, nêu ý kiến, cảm nghĩa của họ bằng cách để lại lời bình luận của mình về bài đăng tin tức đó. |
| 3 | Tìm kiếm tin tức | Người dung nhập thông tin vào thanh tìm kiếm để tìm kiếm tin tức theo tiêu đề và mô tả của bài đăng. |
| 4 | Liên hệ | Khách hàng có thể liên hệ trực tiếp với hệ thống để quảng cáo hoặc những phản hồi khác. |
| 5 | Đăng ký tài người dùng | Khách hàng khi sử dụng hệ thống có thể đăng ký tài khoản trong hệ thống. Khi đăng ký cần điền vào mẫu gồm có tên, email, số điện thoại, địa chỉ, mật khẩu. |
| 6 | Sửa thông tin người dùng | Sau khi đăng nhập vào hệ thống bằng tài khoản của mình, khách hàng có thể sửa thông tin của họ. |

* **Chi tiết chức năng:**
* **Xem tin:**

Người dùng truy cập vào địa chỉ của website trên thanh url chọn vào các page trên site để xem thông tin.

* *Đầu vào*: Thông tin yêu cầu của người dùng về tin tức.
* *Xử lý*: Khi người dùng chọn một tin tức bất kì, thông tin mà người dùng gửi sẽ được xử lý, hệ thống sẽ lấy dữ liệu từ cơ sở dữ liệu, gửi lên cho người dùng.
* *Đầu ra*: Thông tin về tin tức bao gồm tiêu đề, mô tả, hình ảnh và phần nội dụng, …
* **Bình luận:**

Người xem có thể chia sẻ cảm nghĩ của họ về sản phẩm, thông tin bình luận của người dùng sẽ được lưu vào database bao gồm tên id người dùng, id tin tức được bình luận, và nội dung mà người dùng bình luận.

* *Đầu vào*: Thông tin về người bình luận bao gồm tên, ngày giờ và nội dung mà người xem bình luận.
* *Xử lý*: Kiểm tra nội dung mà người dùng nhập vào, nếu hợp lệ thì lưu dữ liệu đó vào database, nếu không thì thông báo lỗi cho người dùng.
* *Đầu ra*: Thông báo từ hệ thống, nếu hợp lệ thì thông báo “Bình luận đã được đăng!”. Nếu không hợp lệ, tức là người dùng không nhập bình luận gì nhưng vẫn nhấn nút gửi thì sẽ có thông báo: “Bạn chưa nhập nội dung”…
* **Tìm kiếm tin tức:**

Để không mất thời gian duyệt từng tin tức của các loại tin trên website, người dùng có thể biết thông tin về tin tức mình muốn xem bằng cách nhập thông tin vào trên thanh tìm kiếm.

* *Đầu vào*: Một từ khoá mà người dùng muốn tìm kiếm.
* *Xử lý*: Khi người dùng nhập từ khoá muốn tìm kiếm, hệ thống sẽ so sánh giá trị đó với dữ liệu có trong database. Nếu nhận được kết quả sẽ trả kết quả tìm kiếm về cho người dùng, nếu không tìm thấy kết quả nào thì hệ thống sẽ gửi thông báo: “Tìm thấy tổng cộng 0 tin tức có liên quan đến từ khóa”.
* *Đầu ra*: Thông báo, kết quả từ hệ thống.
* **Liên hệ:**

Khi khách hàng có các nhu cầu quảng cáo, hoặc một số phát sinh thêm, họ có thể liên hệ trực tiếp qua số điện thoại, địa chỉ email hay đếm trực tiếp địa chỉ đã có trong phần liên hệ của hệ thống.

* *Đầu vào:* Họ tên, email hoặc số điện thoại của người liên hệ, nội dung liên hệ.
* *Xử lý:* Khi khách hàng gửi liên hệ tới hệ thống, quản trị viên sẽ tiếp nhận và xử lý sau đó gọi điện thoại hoặc gửi email tới người liên hệ.
* *Đầu ra:* Kết quả sẽ được gửi tới qua gọi điện trực tiếp hoặc email.
* **Đăng ký thành viên:**

Khi sử dụng hệ thống, người dùng có thể đăng ký làm thành viên của hệ thống để tiện cho việc đánh giá, nêu lên cảm nghĩ của họ.

* + *Đầu vào:* Thông tin người dùng bao gồm: tên, email, mật khẩu.
  + *Xử lý:* Nếu người dùng nhập đúng những thông tin cần thiết, hệ thống sẽ báo “Đăng ký tài khoản thành công!”, nếu sai hệ thống sẽ báo từng lỗi của các thông tin mà người dùng đã nhập sai. Chẳng hạn như: “Bạn chưa nhập đúng định dạng Email!”, “Mật khẩu gồm tối thiểu 6 ký tự!”, …
  + *Đầu ra:* Thông báo từ hệ thống.
* **Sửa thông tin thành viên:**

Sau khi đăng ký làm thành viên, nếu khách hàng muốn thay đổi thông tin của mình thì họ có thể làm điều đó bằng cách truy cập vào thay đổi thông tin tài khoản của mình.

* + *Đầu vào:* Thông tin cần sửa đổi.
  + *Xử lý:* Nếu người dùng nhập đúng, hệ thống sẽ thông báo “Thay Đổi thông tin Người Dùng thành công!”, nếu sai hệ thống sẽ báo từng lỗi của các thông tin mà người dùng đã nhập sai. Chẳng hạn như “Mật khẩu gồm tối thiểu 6 ký tự!”, … Đặc biệt ở phần này sẽ không cho phép người dùng sửa địa chỉ email đã đăng ký.
  + *Đầu ra:* Thông báo từ hệ thống.

#### **3.1.2.2 Back-end**

Từ mô tả, định nghĩa bài toán xây dựng trang web tin tức, ta có thể xác định một số chức năng sẽ có dành cho người quản trị:

* **Các chức năng:**

Bảng 3.2 Bảng chức năng của Back-end

|  |  |  |
| --- | --- | --- |
| **STT** | **Chức năng** | **Mô tả** |
| 1 | Quản lý tin tức | Quản lý mọi thông tin của tin tức. Người quản trị viên có thể thêm, xoá bất kì tin tức nào. Hoặc cập nhật tiêu đề, mô tả, hình ảnh và nội dung của tin tức đó. |
| 2 | Quản lý loại tin | Quản lý loại tin của hệ thống bao gồm các chức năng thêm, xóa, sửa loại tin. |
| 3 | Quản lý slide | Quản lý slide hiển thị trên trang chủ của hệ thống bao gồm các chức năng thêm, xóa, sửa slide. |
| 4 | Quản lý người dùng | Quản lý các thông tin của người dùng khi tạo tài khoản. Phân quyền cho người dùng bình thường và quản trị viên. Thêm mới, sửa, xóa thông tin người dùng. |
| 5 | Xem bình luận | Xem danh sách bình luận của người dùng về các tin tức trong bài đăng. |

* **Chi tiết chức năng**
* **Quản lý tin tức:**

Quản lý thông tin tin tức bao gồm các thuộc tính như: tiêu đề, mô tả, hình ảnh và nội dung, … Có các chức năng như là thêm, xóa, sửa tin tức.

* *Đầu vào:* Thông tin tin tức: tiêu đề, mô tả, hình ảnh và nội dung, …
* *Xử lý*:
* Thêm tin tức: Khi người quản trị hệ thống chọn chức năng thêm tin tức, hệ thống sẽ chuyển đến trang thêm tin tức. Ở trang này người quản trị nhập thông tin về tin tức sau đó gửi thông tin đã nhập cho hệ thống.
* Sửa tin tức: Khi người quản trị hệ thống chọn chức năng sửa tin tức thì thông tin tin tức cần sửa sẽ được hệ thống chuyển đến trang sửa sản phẩm. Ở trang này người quản trị chỉnh sửa thông tin và gửi về cho hệ thống.
* Xóa tin tức: Người quản trị chọn tin tức cần xóa, hệ thống sẽ gửi thông tin tin tức mà người dùng quản trị muốn xóa về database, tin tức sẽ được xóa ở database.
* *Đầu ra*: Danh sách tin tức.
* **Quản lý loại tin:**

Bao gồm các chức năng thêm, sửa xóa loại tin.

* *Đầu vào*: Thông tin các loại tin.
* *Xử lý*: Khi người quản trị website chọn các chức năng thêm, xóa, sửa thì thông tin về loại tin sẽ được cập nhật vào database.
* *Đầu ra*: Danh sách các loại tin.
* **Quản lý slide:**

Bao gồm các chức năng thêm, sửa xóa slide.

* *Đầu vào*: Thông tin danh sách slide đang được hiển thị.
* *Xử lý*: Khi người quản trị website chọn các chức năng thêm, xóa, sửa thì thông tin về các slide sẽ được cập nhật vào database.
* *Đầu ra*: Danh sách slide.
* **Quản lý người dùng:**

Liệt kê tài khoản thành viên của hệ thống. Gồm các chức năng thêm, sửa, xóa thành viên.

* *Đầu vào*: Thông tin người dùng: Tên, email, mật khẩu.
* *Xử lý*:
* Thêm người dùng: Khi người quản trị chọn thêm người dùng, trang quản trị sẽ chuyển tới trang thêm người dùng, người quản trị nhập thông tin và gửi lên hệ thống, nếu hợp lệ (email và tên người dùng là bắt buộc, mật khẩu và mật khẩu nhập lại phải giống nhau) sẽ thêm người dùng vào database, nếu không sẽ báo lỗi. Đồng thời khi thêm sẽ có luôn phần phân quyền. Tài khoản để xem người dùng này có các quyền như người quản trị hay không.

Cụ thể là những quản trị viên là những người mọi quyền hành, sử dụng được tất cả các chức năng trong hệ thống từ Front-end đến Back-end. Còn những người dùng bình thường thì chỉ có các chức năng ở Front-end mà thôi.

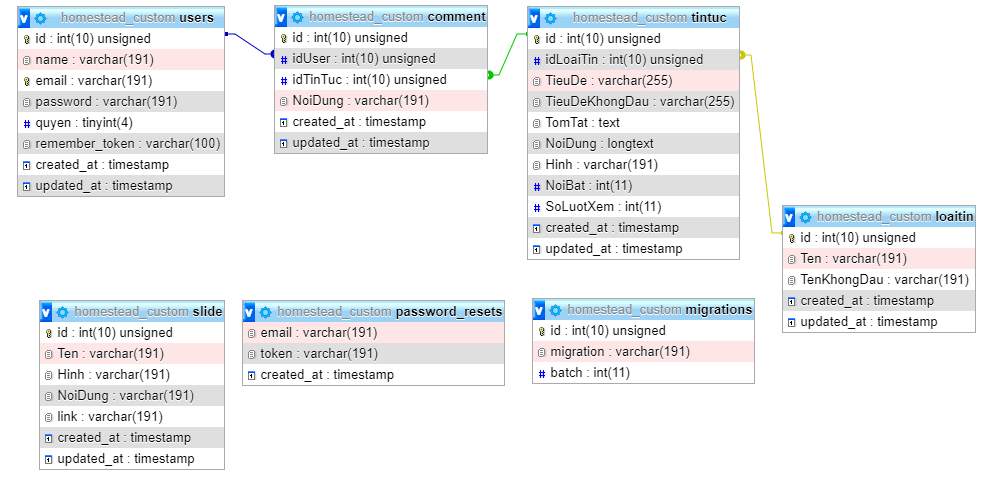
* Sửa người dùng: Khi người quản trị chọn chức năng sửa người dùng thì thông tin người dùng cần sửa sẽ được hệ thống chuyển đến trang sửa người dùng. Ở trang này người quản trị chỉnh sửa thông tin và gửi về cho hệ thống.
* Xóa người dùng: Người quản trị chọn người dùng cần xóa và hệ thống sẽ xóa ở database.
* *Đầu ra*: Danh sách người dùng.
* **Xem bình luận:**

Liệt kê tài khoản thành viên của hệ thống. Gồm các chức năng thêm, sửa, xóa thành viên.

* *Đầu vào*: Thông tin người dùng: Họ tên, số điện thoại, mail, địa chỉ khách hàng, mật khẩu.
* *Xử lý*:
* Thêm người dùng: Khi người quản trị chọn thêm người dùng, trang quản trị sẽ chuyển tới trang thêm người dùng, người quản trị nhập thông tin và gửi lên hệ thống, nếu hợp lệ (email và tên người dùng là bắt buộc, mật khẩu và mật khẩu nhập lại phải giống nhau) sẽ thêm người dùng vào database, nếu không sẽ báo lỗi.
* Sửa người dùng: Khi người quản trị chọn chức năng sửa người dùng thì thông tin người dùng cần sửa sẽ được hệ thống chuyển đến trang sửa người dùng. Ở trang này người quản trị chỉnh sửa thông tin và gửi về cho hệ thống.
* Xóa người dùng: Người quản trị chọn người dùng cần xóa và hệ thống sẽ xóa ở database.
* *Đầu ra*: Danh sách người dùng.

### **3.1.3 Cơ sở dữ liệu**

#### **3.1.3.1 Các bảng trong cơ sở dữ liệu**



Hình 3.1 Các bảng trong cơ sở dữ liệu

#### **3.1.3.2 Chi tiết các bảng dữ liệu**

Cơ sở dữ liệu có tổng cộng 5 bảng dữ liệu sau:

Bảng 3.3. Chi tiết bảng loại tin

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Diễn giải** |
| 1 | ID | INT(10) | Mã loại tin. Là khóa chính của bảng. |
| 2 | TEN | VARCHAR(191) | Tên danh mục sản phẩm. |
| 3 | TENKHONGDAU | VARCHAR(191) | Dựa vào tên của loại tin để làm đường dẫn cho url. |
| 4 | CREATED\_AT | TIMESTAMP | Thời gian nhập loại tin. |
| 5 | UPDATED\_AT | TIMESTAMP | Thời gian cập loại tin. |

Bảng 3.4. Chi tiết bảng tin tức

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Diễn giải** |
| 1 | ID | INT(10) | Mã tin tức. Là khóa chính của bảng. |
| 2 | IDLOAITIN | INT(10) | Mã loại tin của tin tức. Là khóa ngoại liên kết tới trường ID của bảng LoaiTin. |
| 3 | TIEUDE | VARCHAR(255) | Tiêu đề của tin tức. |
| 4 | TIEUDEKHONGDAU | VARCHAR(255) | Dựa vào tiêu đề của tin tức để làm đường dẫn cho url. |
| 5 | TOMTAT | TEXT | Mô tả của tin tức. |
| 6 | NOIDUNG | LONGTEXT | Nội dung của tin tức. |
| 7 | HINH | VARCHAR(191) | Hình của tin tức. |
| 8 | NOIBAT | INT(11) | Phân biệt đây có phải là tin tức nổi bật hay không. |
| 9 | SOLUOTXEM | INT(11) | Chứa số lượt xem của tin tức. |
| 10 | CREATED\_AT | TIMESTAMP | Thời gian nhập tin tức. |
| 11 | UPDATED\_AT | TIMESTAMP | Thời gian cập nhật tin tức. |

Bảng 3.5. Chi tiết bảng comment

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Diễn giải** |
| 1 | ID | INT(10) | Mã bình luận. Là khóa chính của bảng. |
| 2 | IDUSER | INT(10) | Mã người dùng. Là khóa ngoại liên kết tới trường ID của bảng Users. |
| 3 | IDTINTUC | INT(10) | Mã tin tức. Là khóa ngoại liên kết tới trường ID của bảng TinTuc. |
| 4 | NOIDUNG | VARCHAR(191) | Nội dung phần bình luận của người dùng về tin tức. |
| 5 | CREATED\_AT | TIMESTAMP | Ngày tạo bình luận. |
| 6 | UPDATED\_AT | TIMESTAMP | Ngày cập nhật bình luận. |

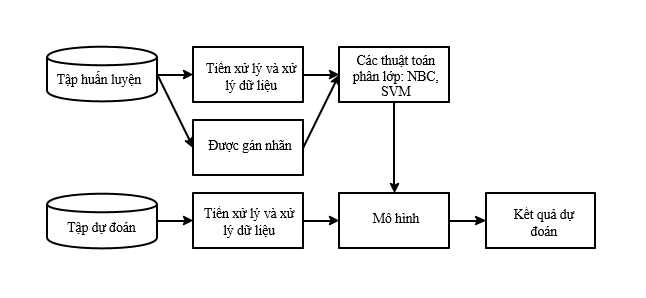
Bảng 3.6. Chi tiết bảng users

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Diễn giải** |
| 1 | ID | INT(10) | Mã người dùng hoặc quản trị viên. Là khóa chính của bảng. |
| 2 | NAME | VARCHAR(191) | Tên của người dùng hoặc quản trị viên. |
| 3 | EMAIL | VARCHAR(191) | Email đăng nhập cho dùng người hoặc quản trị viên. |
| 4 | PASSWORD | VARCHAR(191) | Mật khẩu đăng nhập cho người dùng hoặc quản trị viên. |
| 5 | QUYEN | TINYINT(4) | Phân quyền tài khoản giữa người dùng bình thường và quản trị viên. |
| 6 | REMEMBER\_TOKEN | VARCHAR(100) | Lưu trữ token cho tài khoản đó. |
| 7 | CREATED\_AT | TIMESTAMP | Ngày lập tài khoản người dùng hoặc quản trị viên. |
| 8 | UPDATED\_AT | TIMESTAMP | Ngày cập nhật tài khoản người dùng hoặc quản trị viên. |

Bảng 3.7. Chi tiết bảng slide

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Diễn giải** |
| 1 | ID | INT(11) | Mã slide. Là khóa chính của bảng. |
|  | TEN | VARCHAR(191) | Tên slide ảnh. |
| 2 | HINH | VARCHAR(191) | Hình ảnh của các Slide hiển thị trong trang chủ. |
| 3 | NOIDUNG | VARCHAR(191) | Nội dung của slide ảnh. |
| 4 | LINK | VARCHAR(191) | Link của slide ảnh. |
| 5 | CREATED\_AT | TIMESTAMP | Ngày tạo slide. |
| 6 | UPDATED\_AT | TIMESTAMP | Ngày cập nhật slide. |

## **3.2 Ứng dụng thuật toán phân loại Naive Bayes Classifier**

 Hình 3.2 Mô hình phân lớp văn bản

Để có thể hoàn thành được mô hình phân loại văn bản thì cần khá nhiều bước khác nhau mới có thể đưa ra kết quả dự đoán. Dựa vào mô hình trên, có thể chia thành hai quá trình chính như sau:

* Quá trình thứ nhất:
* Thu thập dữ liệu cho tập huấn luyện.
* Tiền xử lý dữ liệu.
* Gán nhãn.
* Lựa chọn thuật toán phù hợp.
* Đưa ra mô hình.
* Quá trình thứ hai:
* Thu thập dữ liệu cho tập dự đoán.
* Tiền xử lý dữ liệu.
* Đưa vào mô hình đã có ở trên.
* Cuối cùng xuất ra được kết quả dự đoán.

### **3.2.1 Tiến hành thu thập dữ liệu:**

Thu thập dữ liệu hay còn được gọi là kỹ thuật crawl dữ liệu website, được sử dụng trong thực tế rất nhiều để thu thập thông tin, dữ liệu trên website. Để tạo một web crawl có rất nhiều cách và nhiều ngôn ngữ hỗ trợ như Python, Javascript, Java, PHP, … Nhưng ở đây tôi sẽ chỉ dùng Javascript trên môi trường NodeJS, cụ thể là thư viện Puppeteer để thực hiện crawling dữ liệu.

Để chuẩn bị cho việc tiến hành thu thập dữ liệu, tôi đã cài đặt NodeJS và NPM (Node Package Manager) trong máy của mình. Sau khi đã có NodeJS và NPM trong máy thì tôi tiếp tục cài đặt thư viện Puppeteer trong thư mục của mình thông qua câu lệnh.

***npm install --save puppeteer***

Sau khi đã có Puppeteer trong thư mục, tôi tiến hành code để thu thập dữ liệu và đây là đoạn code mẫu để thu thập thông tin url của 10 tin tức đầu tiên trong trang web “https://tuoitre.vn/xe.htm”:

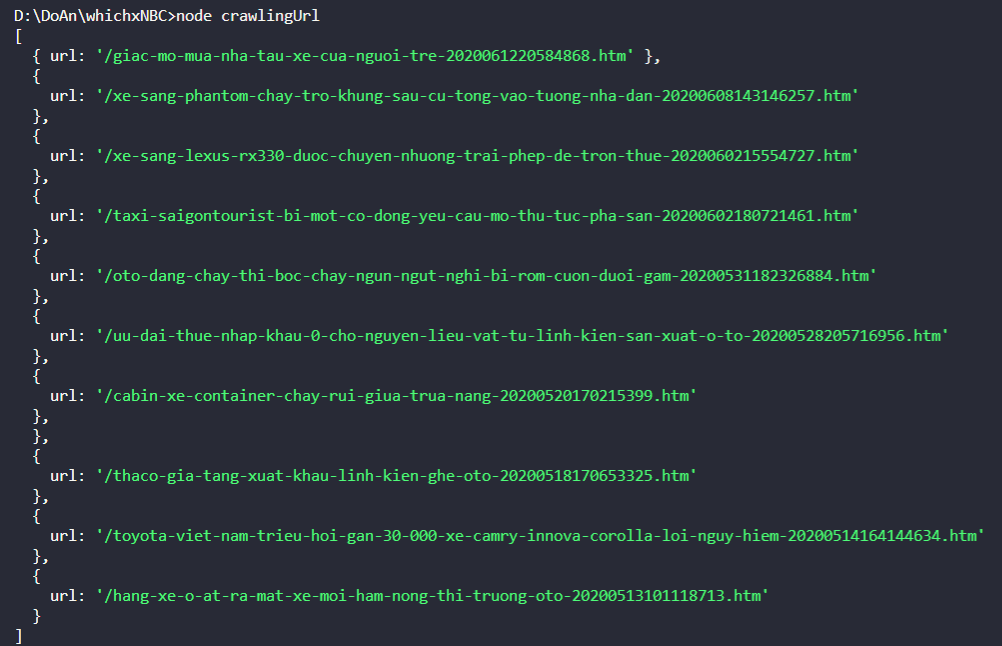


Hình 3.3 Đoạn code thu thập dữ liệu url của một trang web

Giải thích một chút về đoạn code trên, trong đoạn code này sử dụng khá nhiều API (Application Programming Interface) như:

* **puppeteer.launch:** Mở trình duyệt Chrome lên để bắt đầu làm trò. Hàm này trả về object kiểu Browser.
* **browser.newPage:** Mở một tab mới trong Chrome để làm trò. Hàm này trả về object kiểu Page.
* **page.goto:** Đi tới một trang nào đó. Có params waitUntil khá quan trọng. Params này quyết định chúng ta chờ tới khi page vừa mới load xong, hay sau khi page đã load toàn bộ JavaScript và hình ảnh.
* **page.evaluate:** Đây là API quan trọng nhất, cho phép ta chạy script trong browser và lấy kết quả trả về. Chúng ta sẽ dùng API này để cào mương 14 nhé.
* **browser.close:** Tắt trình duyệt (Đỡ phải tắt bằng tay)

Và đây là kết quả sau khi chạy file trên:

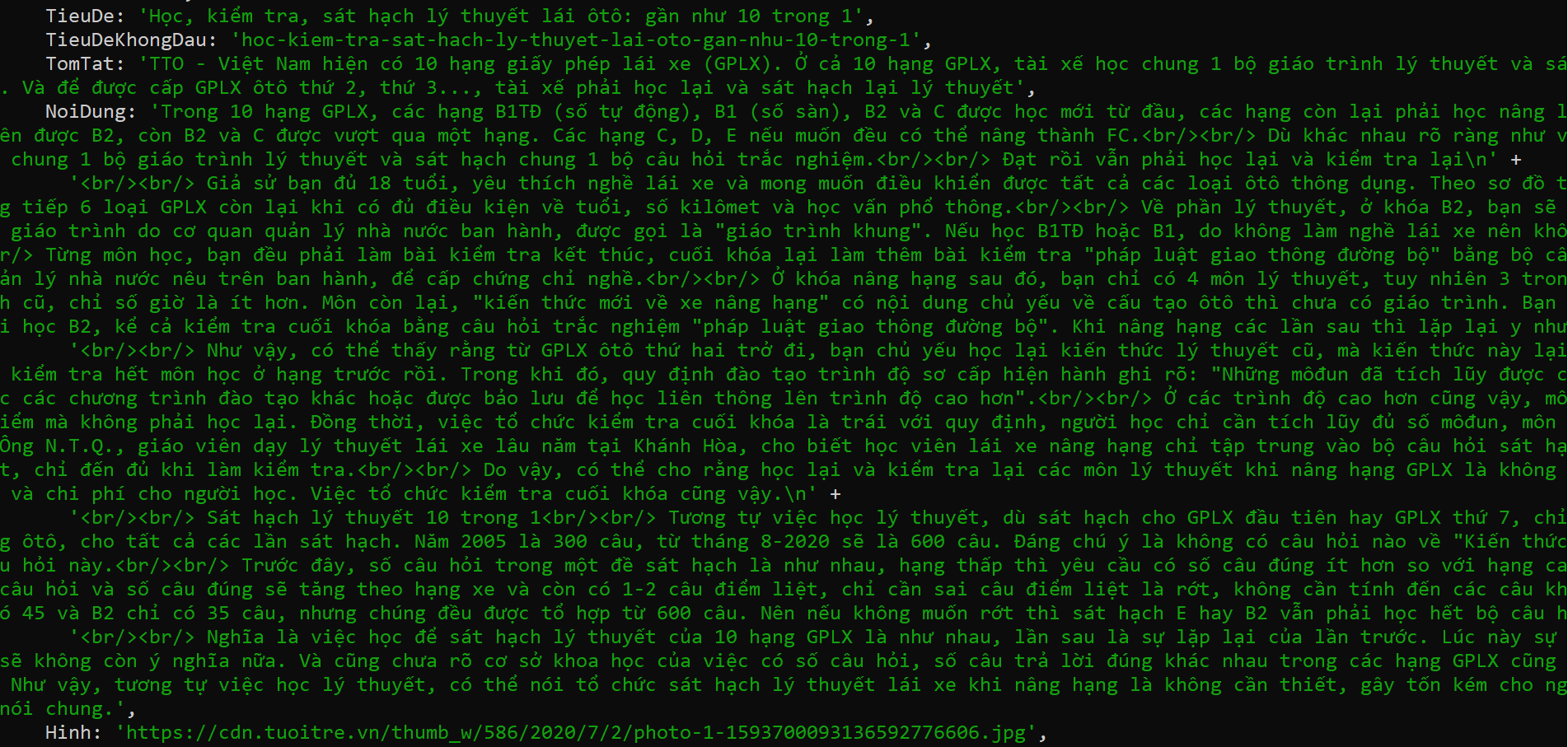


Hình 3.4 Kết quả trả về sau khi thu thập dữ liệu url

Có đúng 10 url được crawling về chứa trong từng object nằm trong cùng một mảng.

Trên đây là kỹ thuật crawling dữ liệu của một website, giúp tiết kiệm được thời gian cũng như công sức trong việc lấy dữ liệu của một trang web. Tương tự đó tôi đã tự nghiên cứu và mở rộng ra mô hình lớn hơn để thu thập các dữ liệu khác cần cho hệ thống như tiêu đề, mô tả, hình ảnh hay nội dung của một tin tức.

Và đây là thành quả khi lấy dữ liệu của một tin tức thuộc thể loại Xe.



Hình 3.5 Kết quả trả về khi thu thập các thông tin cần thiết của một tin tức

### **3.2.2 Phân tích và tiền xử lý dữ liệu trong hệ thống:**

Sau khi đã có được các dữ liệu cần thiết, tôi bắt đầu phân tích và quyết định lấy tiêu đề của tin tức làm dữ liệu cho quá trình phân loại văn bản. Và bây giờ sẽ là quá trình tiền xử lý dữ liệu.

Các bước tiền xử lý dữ liệu bao gồm:

* Xoá bỏ dấu câu (remove punctuation): đây là bước đầu tiên trong quá trình tiền xử lý dữ liệu và nó cũng rất quan trọng vì nó loại bỏ được tất cả các dấu câu trong dữ liệu, làm cho dữ liệu chỉ còn là chữ số.

Ví dụ có một tin tức có tiêu đề là:

“*Xe sang Lexus RX330 được chuyển nhượng trái phép để trốn thuế?*”

Tiêu đề sẽ được xoá hết các dấu câu và thành như thế này:

“*Xe sang Lexus RX330 được chuyển nhượng trái phép để trốn thuế*”

* Tách từ (word segmentation): đây là bước quan trọng bậc nhất trong quá trình tiền xử lý. Vì nếu tách từ sai thì các từ đó trong tập dữ liệu sẽ mang những nghĩa khác nhau, đặc biệt đối với Tiếng Việt thì càng dể nhầm lẫn. Dẫn đến làm sai lệch trong quá trình huấn luyện cũng như dự đoán dữ liệu, từ đó làm cho độ chính xác khi phân loại giảm đi rõ rệt.

Tiếp tục với ví dụ trên ta sẽ được một chuỗi với nội dung:

“*Xe sang Lexus RX330 được chuyển\_nhượng trái\_phép để trốn\_thuế*”

* Chuẩn hoá từ (standardized words): mục đích của chuẩn hoá từ là làm cho các từ trở nên đồng nhất với nhau. Cụ thể ở đây là biến tất cả chữ số trong dữ liệu thành chữ in thường cho đồng bộ với nhau.

Tiếp tục với ví dụ trên ta sẽ được một chuỗi với nội dung:

“*xe sang lexus rx330 được chuyển\_nhượng trái\_phép để trốn\_thuế*”

* Loại bỏ stopwords (remove stopwords): stopwords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở Tiếng Việt stopwords là những từ như: để, này, kia, là, thì, mà, ... Để loại bỏ stopwords, ta tiến hành lọc những từ stopwords xuất hiện trong dữ liệu và loại bỏ chúng đi.

Tiếp tục với ví dụ trên ta sẽ được một chuỗi với nội dung:

“*xe sang lexus rx330 chuyển\_nhượng trái\_phép trốn\_thuế*”

Trên đây là 4 bước trong quá trình tiền xử lý dữ liệu của hệ thống, đây là quá trình cực kì quan trọng vì nó là tiền đề giúp cho bộ huấn luyện và bộ dự đoán trở nên tốt hơn rất nhiều. Từ đó nâng cao độ chính xác khi tiến hành phân loại văn bản.

### **3.2.3 Tiến hành quá trình phân loại dữ liệu:**

Khi đã có được các bước trong quá trình tiền xử lý dữ liệu, ta sẽ tiến hành quá trình phân loại tin tức bản qua từng bước sau:

* Gắn nhãn (labels): tức là điền thông tin các thể loại của tin tức vào cùng một labels (một mảng). Mục đích là để đưa dữ liệu của bộ huấn luyện vào từng nhãn (thể loại). Ngoài ra đây còn là mục tiêu hướng tới trong quá trình dự đoán, tức là các tin tức sau khi phân loại sẽ được gắn thể loại và thuộc vào thể loại đó.
* Chuẩn bị bộ dữ liệu huấn luyện: tiến hành thu thập tin tức từ một website và tiền xử lý chúng trước khi đưa vào bộ huấn luyện.
* Chuẩn bị bộ dữ liệu dự đoán: tiến hành thu thập tin tức từ một website và tiền xử lý chúng trước khi đưa vào bộ dự đoán.
* Tiến hành phân loại: Sau khi đã có đầy đủ hai bộ dữ liệu huấn luyện và dự đoán cùng với các thể loại trong labels, ta bắt đầu tiến hành phân loại tin tức bằng cách:
* Gắn từng bộ dữ liệu huấn luyện vào từng thể loại.
* Thực hiện quá trình phân loại cho bộ dữ liệu dự đoán, kết quả trả về là từng tin tức sẽ được gắn cho một thể loại. Kết quả này phụ thuộc rất nhiều vào dữ liệu của bộ huấn luyện được gắn ở từng thể loại.

Sau khi đã hoàn tất quá trình phân loại tin tức, dữ liệu sẽ được lưu vào cơ sở dữ liệu để được hiển thị lên website.

### **3.3.4 Kết quả:**

Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu.

Sau khi ứng dụng thuật toán phân loại Naive Bayes Classifier vào hệ thống cho kết quả về độ chính xác khá tương đối và đa dạng, rơi vào tầm 78 - 82%. Nhận thấy độ chính xác của thuật toán phụ thuộc rất nhiều vào bộ dữ liệu huấn luyện. Bộ dữ liệu huấn luyện nếu được huấn luyện tốt, số lượng không được quá ít, cũng không nên quá nhiều, chỉ cần một lượng dữ liệu vừa và đủ là có thể mang lại kết quả phân loại tốt.

### **3.2.5 Đánh giá:**

* **Ưu điểm:** Dễ sử dụng và nhanh khi cần đoán nhãn của dữ liệu dự đoán.
* **Nhược điểm:** Độ chính xác của thuật toán phân lớp Naive Bayes Classifier nếu so với các thuật toán khác thì không được cao.
* **Cách khắc phục:**

Khi tiến hành phân loại dữ liệu nhưng lại không đạt được kết quả như mong đợi, cụ thể là độ chính xác thấp hoặc không cao. Có thể tăng độ chính xác cho dự đoán của thuật toán Naive Bayes Classifier bằng cách:

* Thay đổi các bước tiền xử lý dữ liệu. Tuỳ thuộc vào nhu cầu và mục đích của việc phân loại, có thể thêm, bớt hoặc thay đổi vị trí của các bước trong quá trình tiền xử lý dữ liệu.
* Thay đổi bộ dữ liệu huấn luyện. Khi kết quả mang lại không như mong đợi, có thể do bộ huấn luyện đang sử dụng có số lượng quá ít hoặc quá nhiều, hoặc cũng có thể là do không phù hợp.
* Sửa lại mô hình.
* Thay thế mô hình khác mới hơn

# CHƯƠNG 4. KẾT QUẢ THỰC HIỆN ĐỀ TÀI

Sau thời gian nghiên cứu và thực hiện lập trình và cài đặt, dưới đây là các hình ảnh về hệ thống trang web tin tức của tôi.

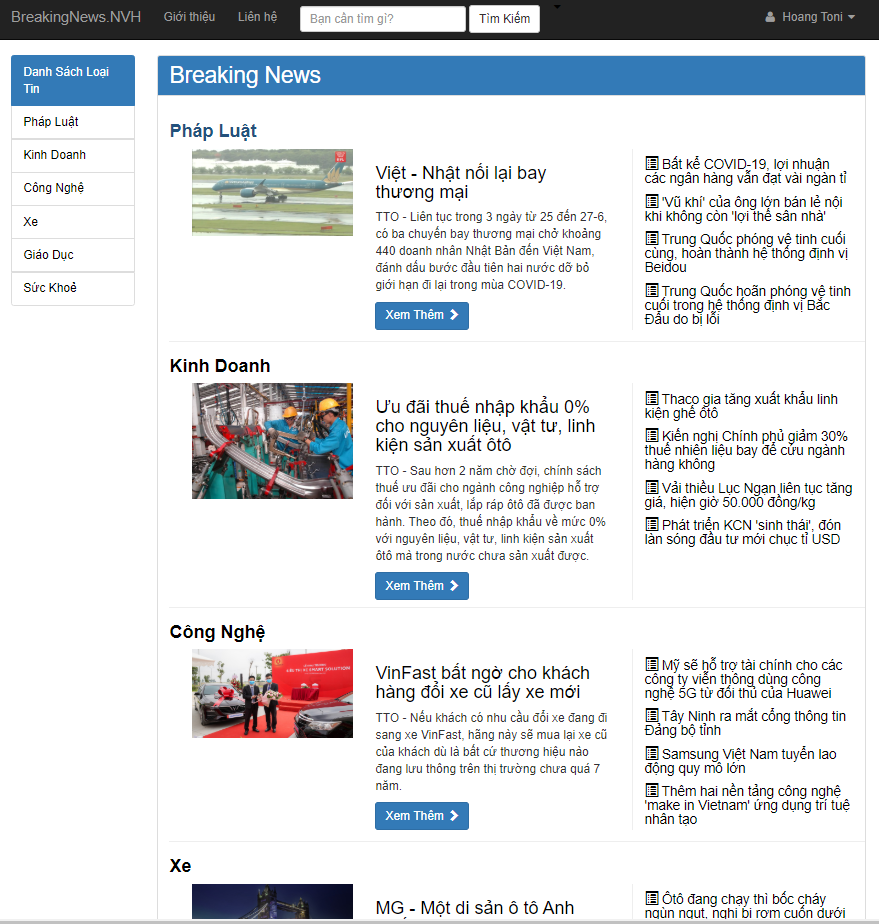
## **4.1 Giao diện người dùng**

* Trang chủ:



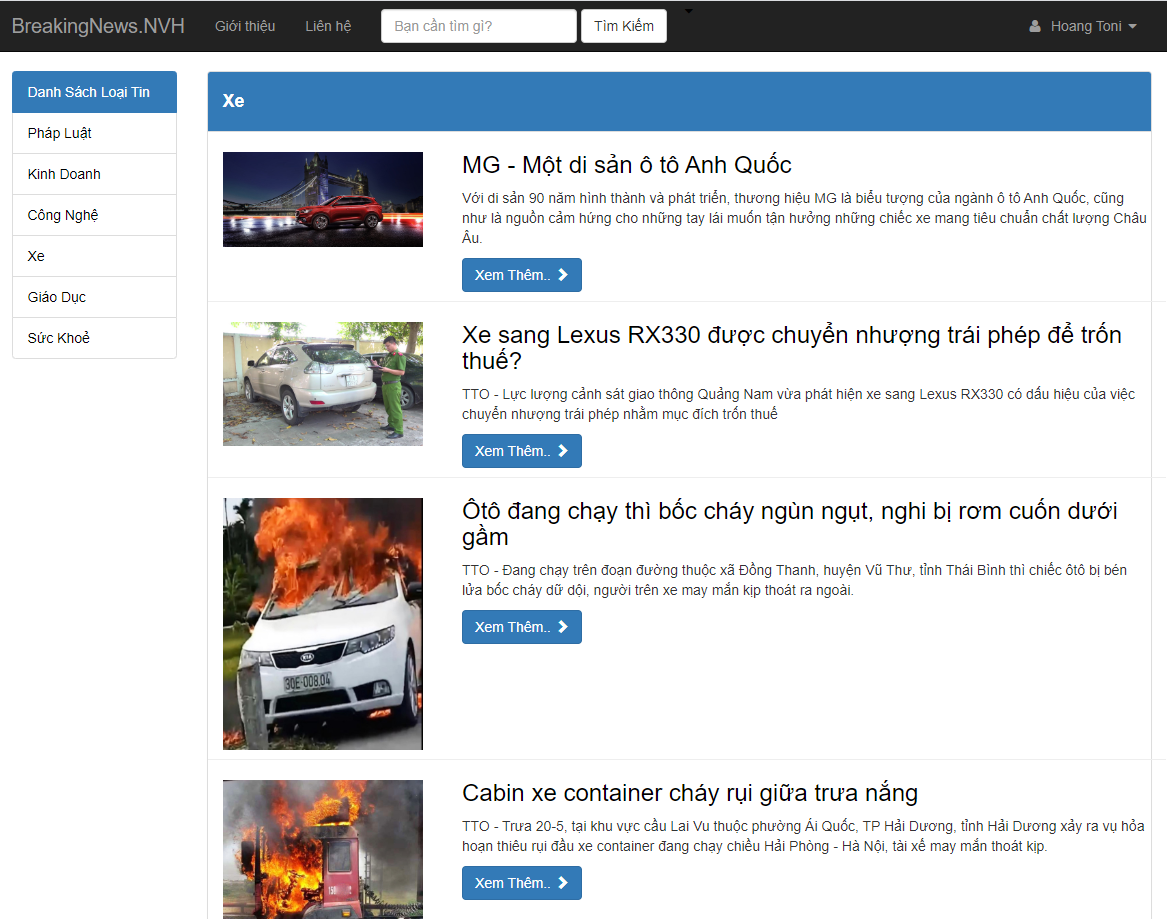
Hình 4.1 Trang chủ

* Giao diện phía dưới của trang chủ



Hình 4.2 Giao diện phía dưới cỉa trang chủ

* Trang danh sách tin tức theo từng loại tin :



Hình 4.3 Trang tin tức theo loại tin Xe

* Trang chi tiết tin tức



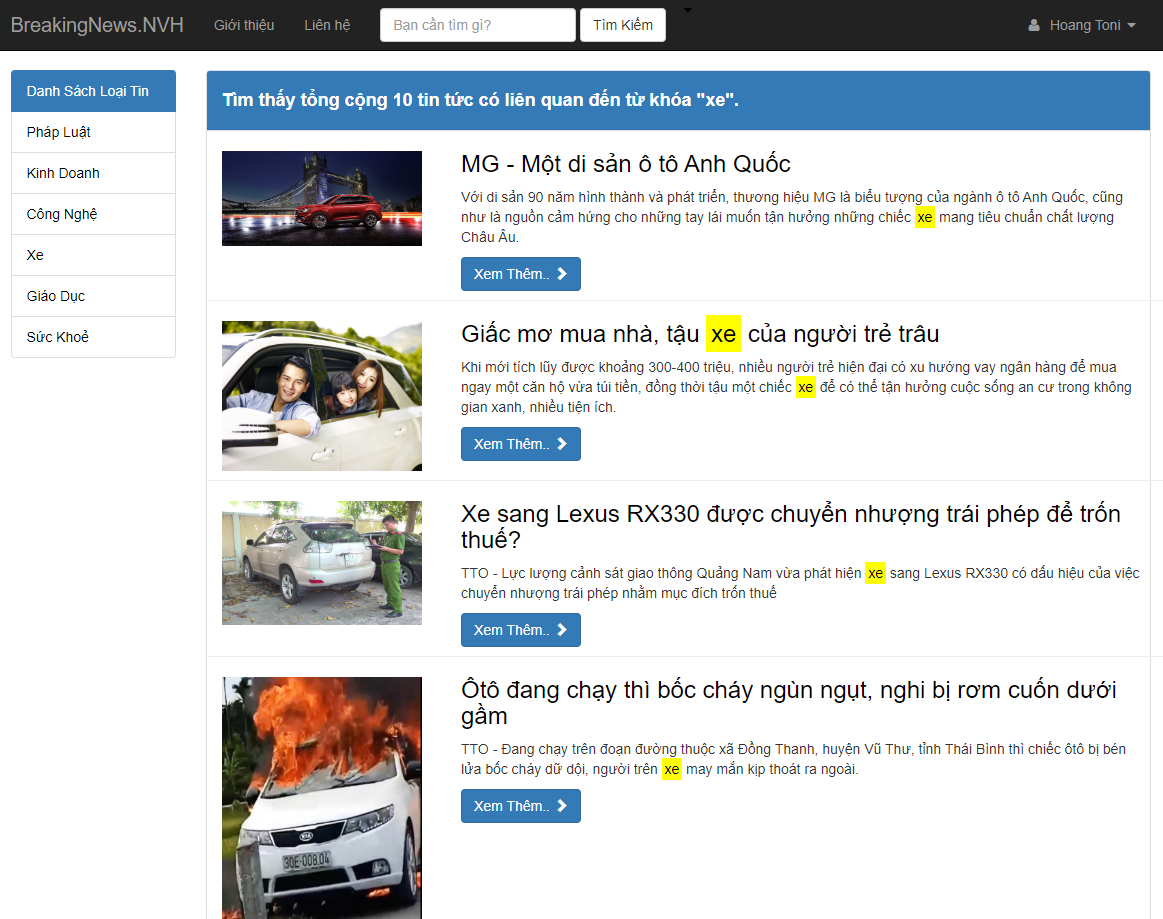
Hình 4.4 Trang chi tiết sản phẩm

* Bên phải trang chi tiết tin tức còn hiển thị giao diện các tin tức cùng loại tin.



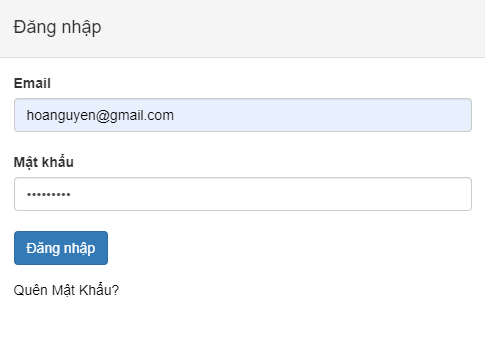
Hình 4.5 Trang chi tiết sản phẩm

* Trang tìm kiếm :



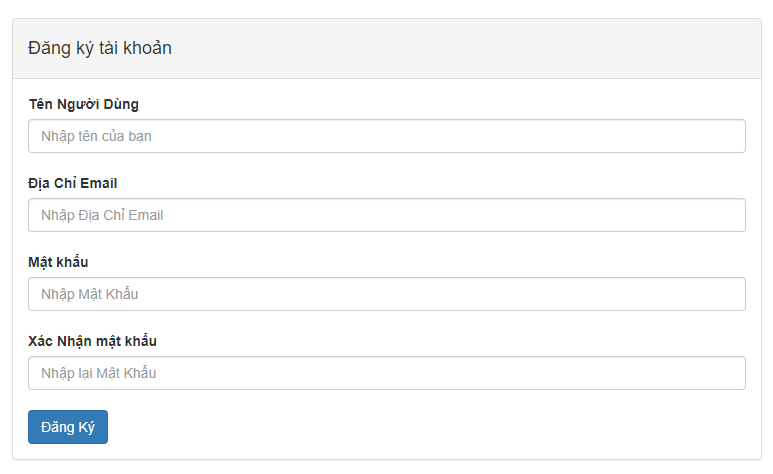
Hình 4.6 Trang đăng nhập

* Giao diện đăng nhập



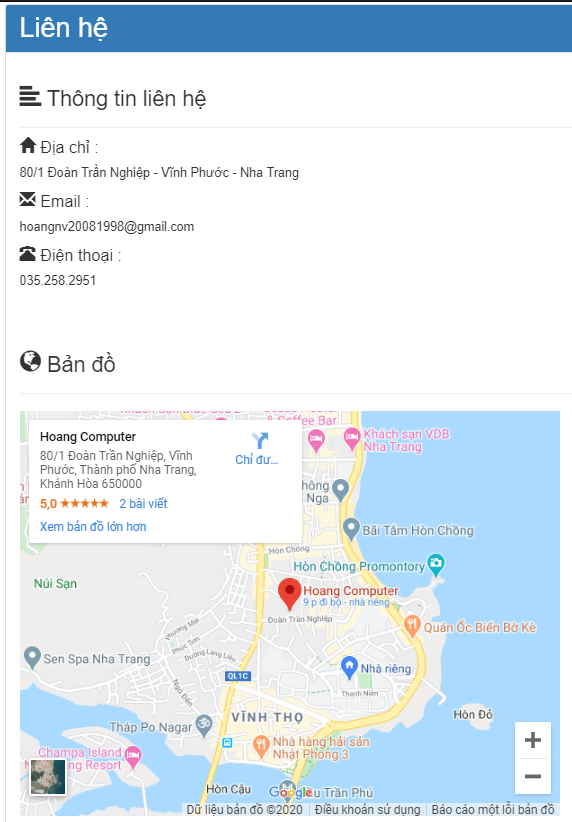
Hình 4.7 Trang đăng nhập

* Trang đăng ký tài khoản



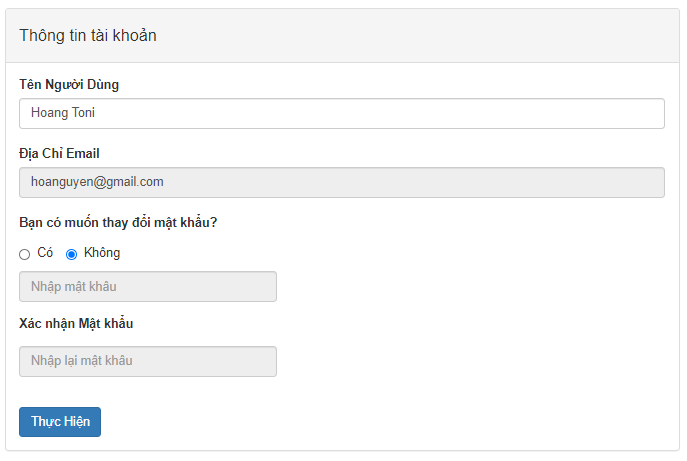
Hình 4.8 Trang đăng ký

* Trang liên hệ



Hình 4.9 Trang liên hệ

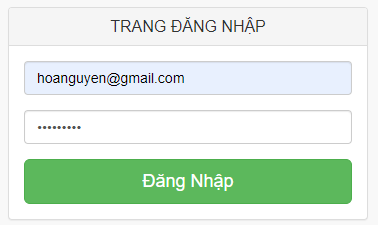
* Trang thiết lập tài khoản:



Hình 4.10 Trang thiết lập tài khoản

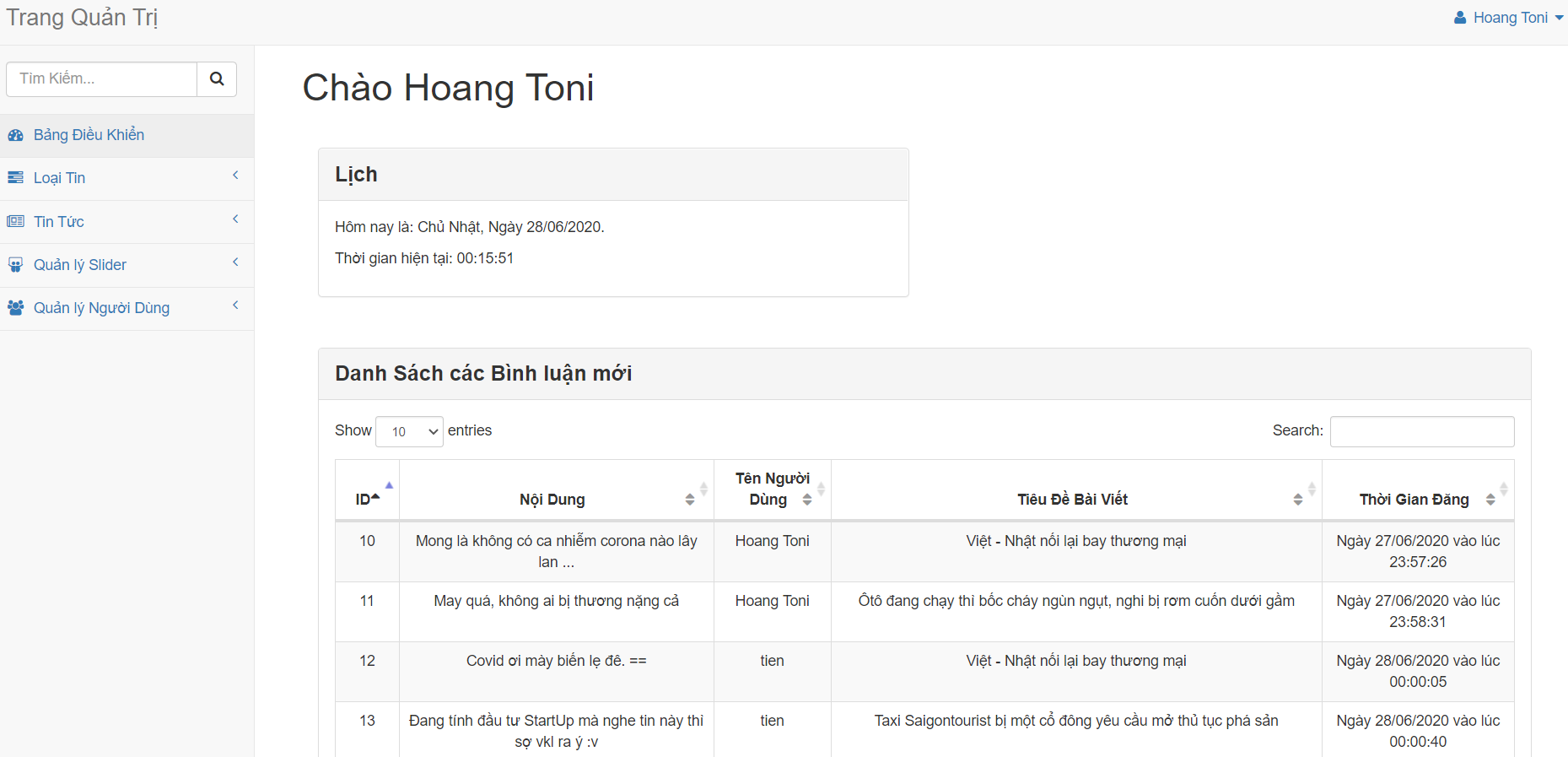
## **4.2 Giao diện ban quản trị**

* Trang đăng nhập của admin



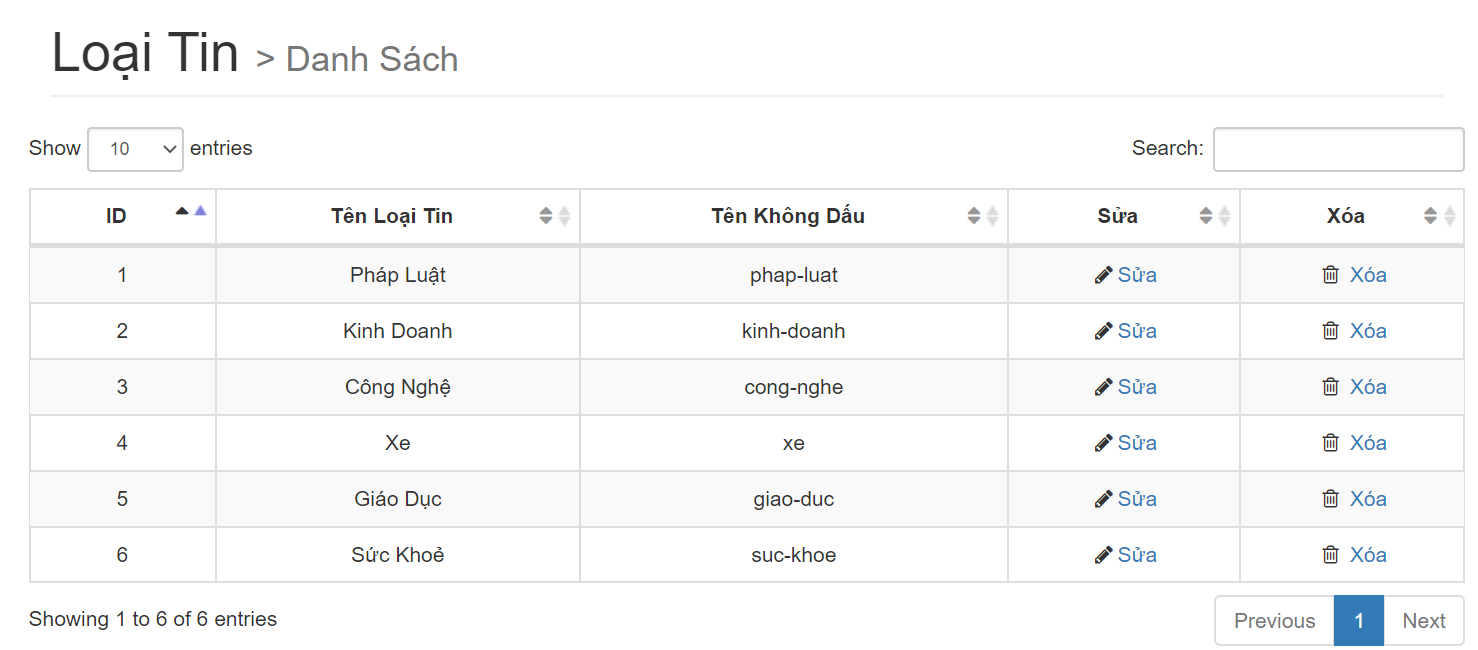
Hình 4.11 Trang login của admin

* Trang Dashboard



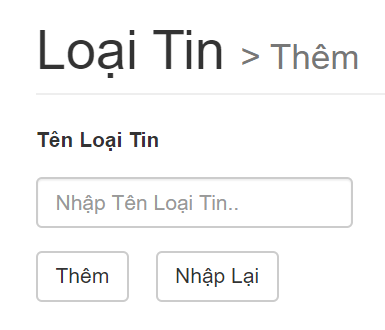
Hình 4.12 Trang dashboard của admin

* Trang quản lý loại tin



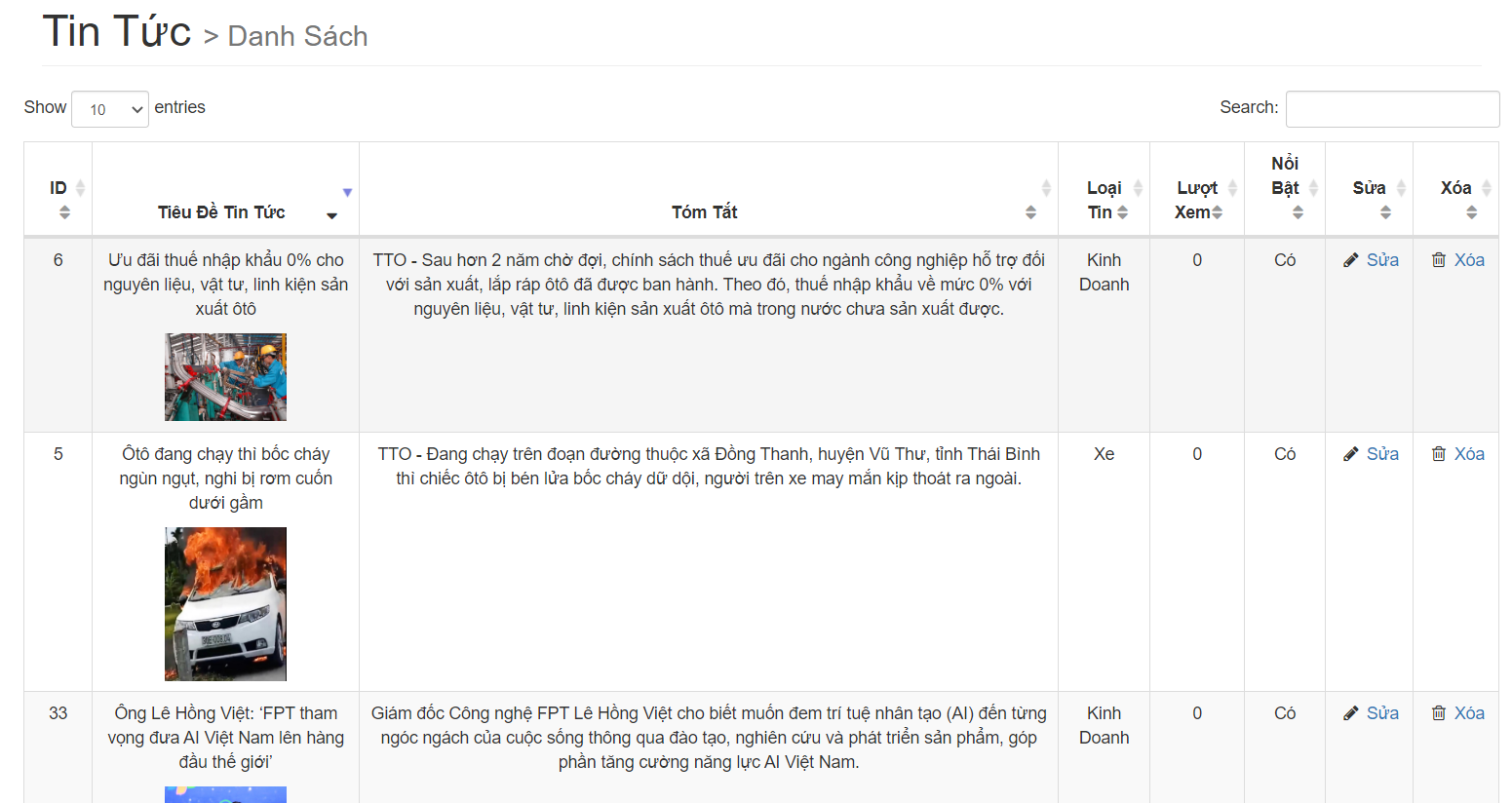
Hình 4.13 Trang quản lý loại tin

* Trang thêm loại tin



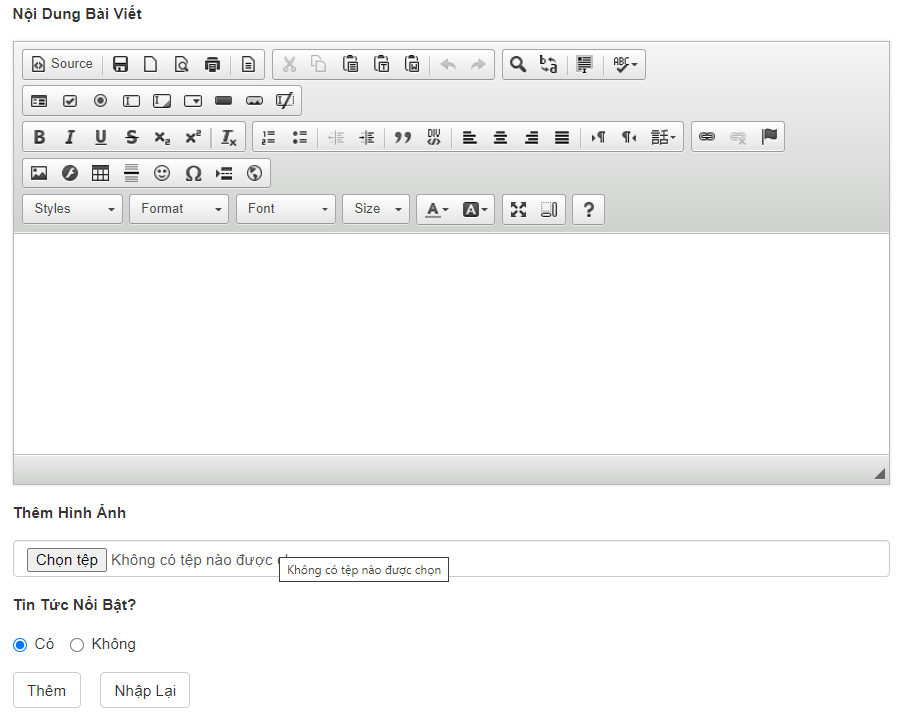
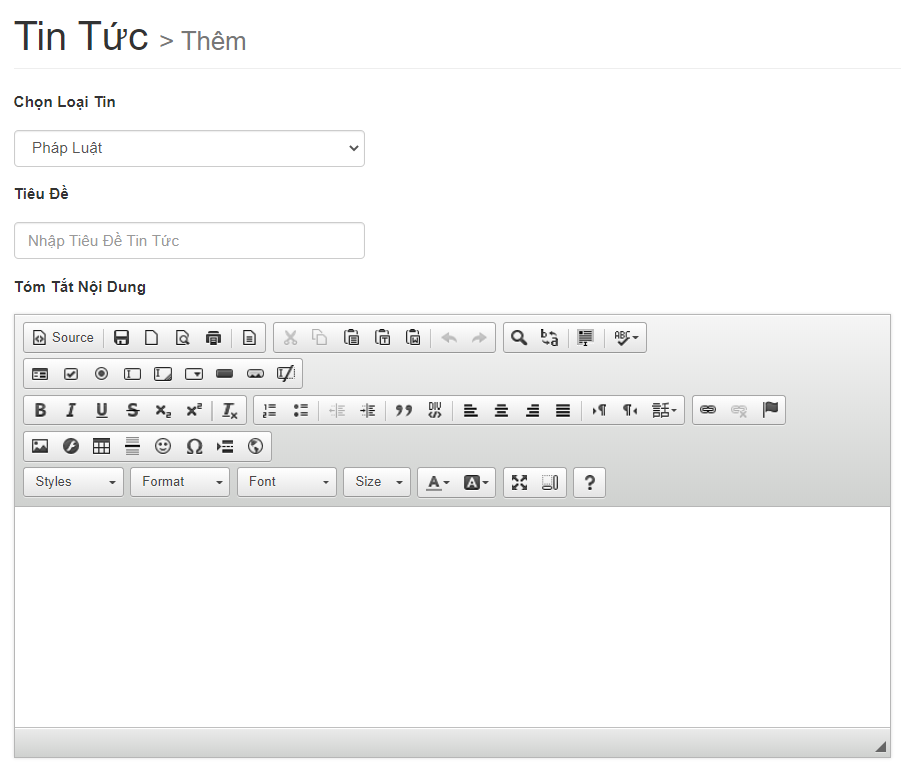
Hình 4.14 Trang thêm loại tin

* Trang quản lý tin tức



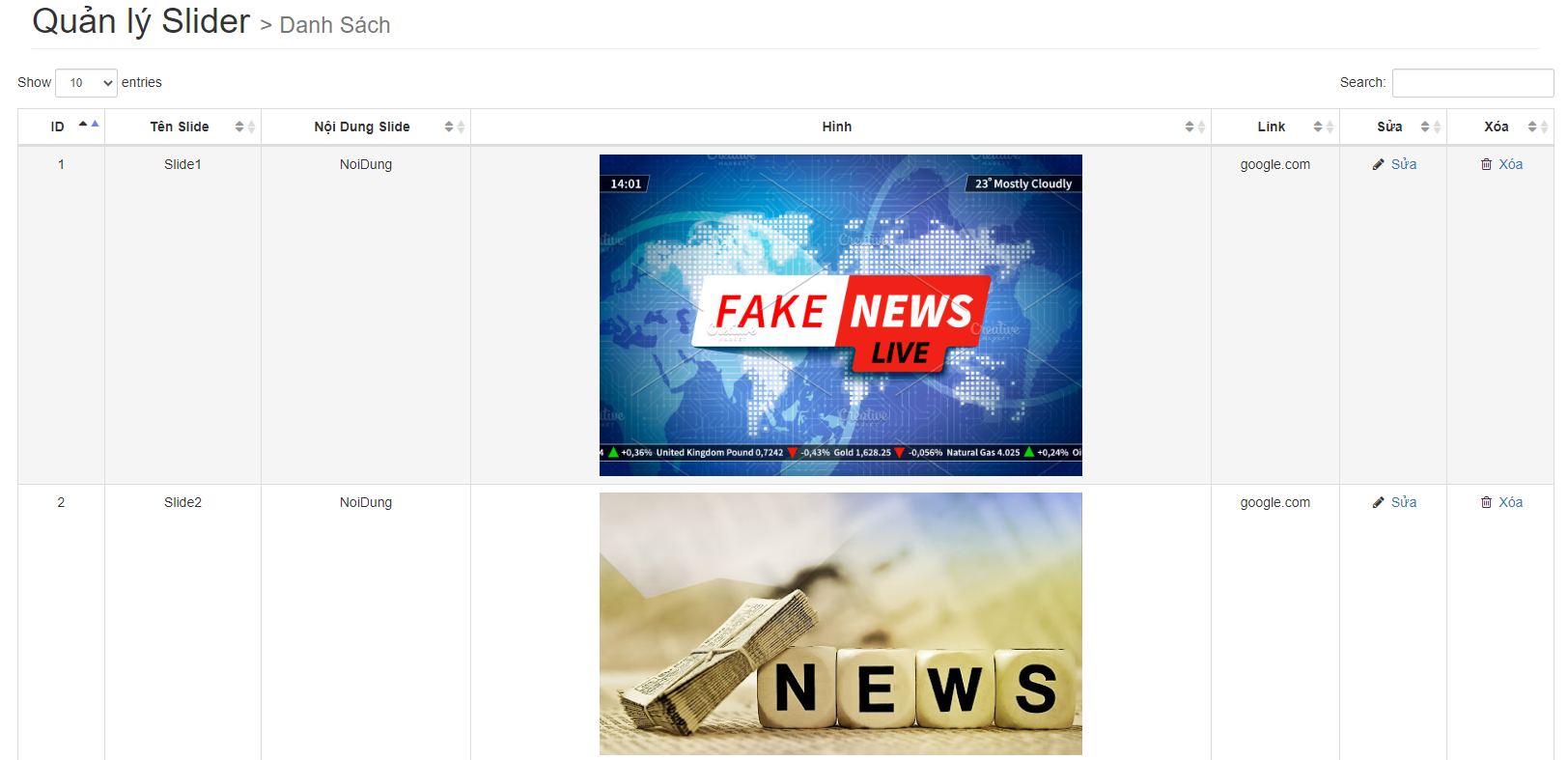
Hình 4.15 Trang quản lý tin tức

* Trang thêm tin tức



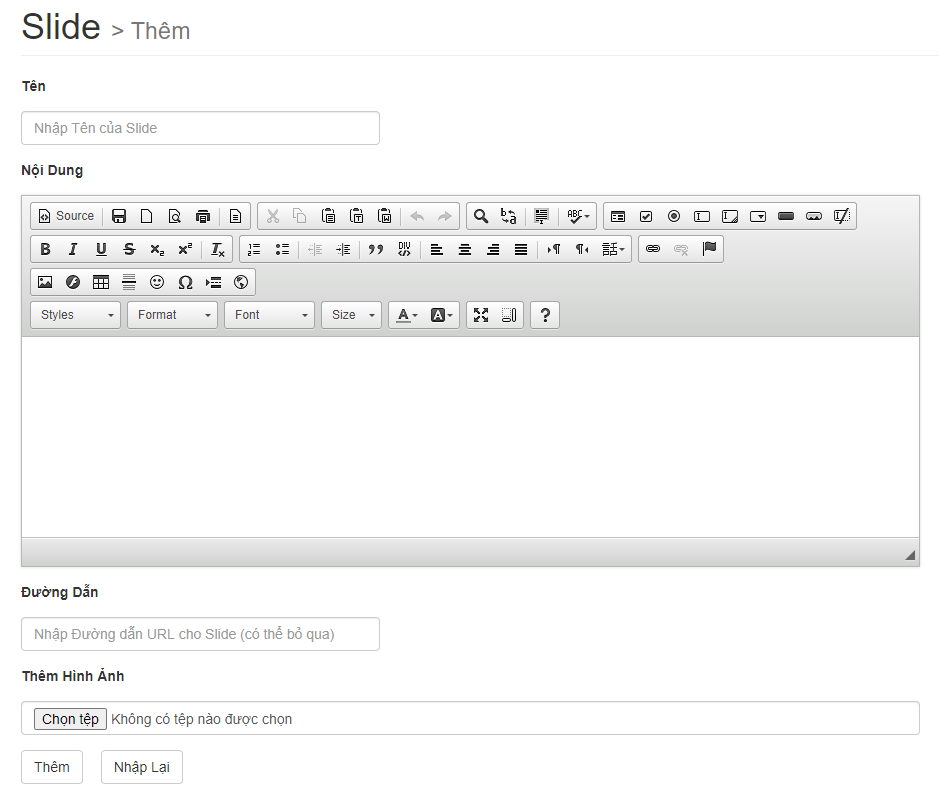
Hình 4.16 Trang thêm tin tức

* Trang quản lý slide



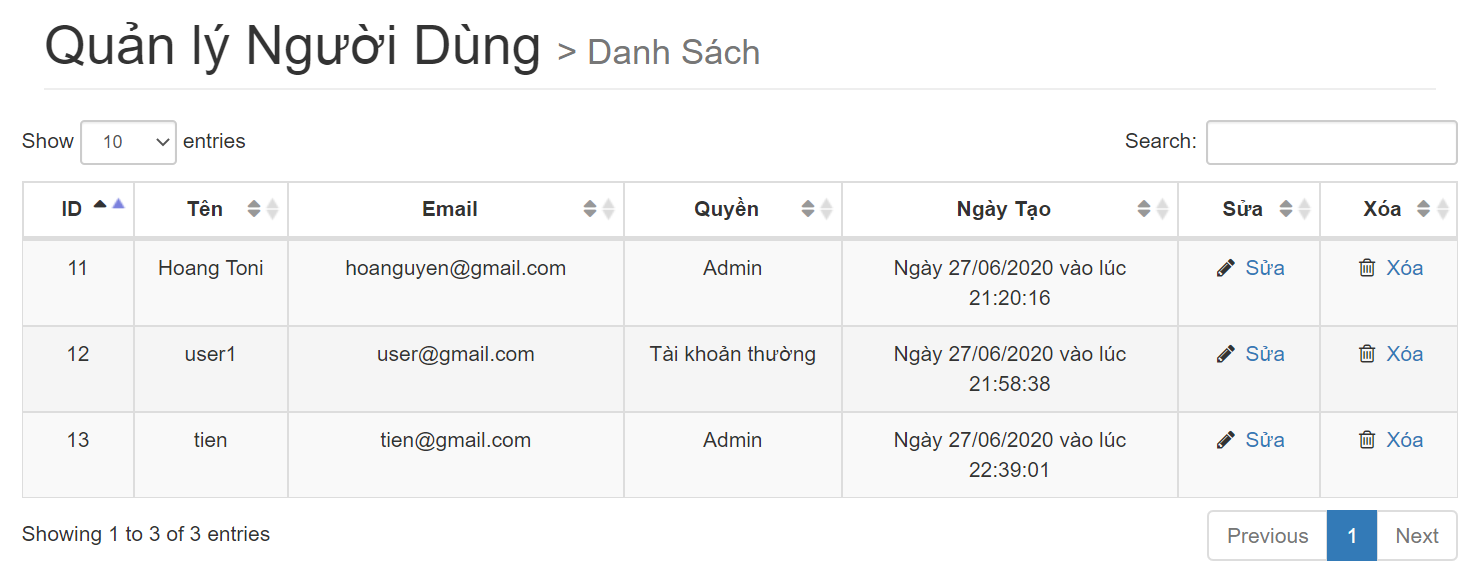
Hình 4.17 Trang quản lý slide

* Trang thêm slide



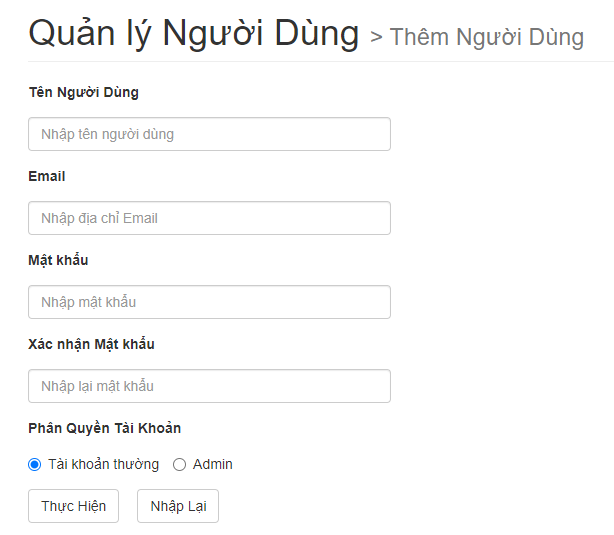
Hình 4.18 Trang thêm slide

* Trang quản lý người dùng



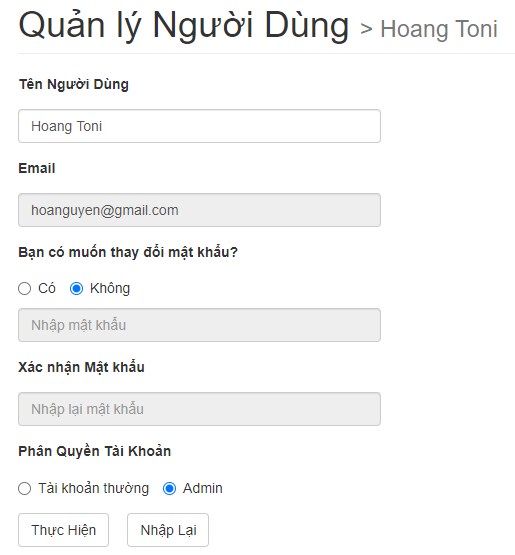
Hình 4.19 Trang quản lý người dùng

* Trang thêm người dùng



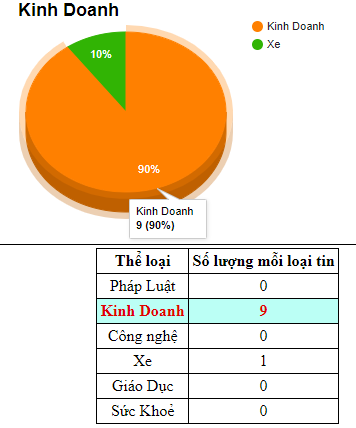
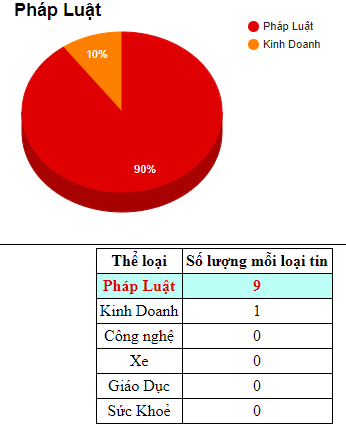
Hình 4.20 Trang thêm người dùng

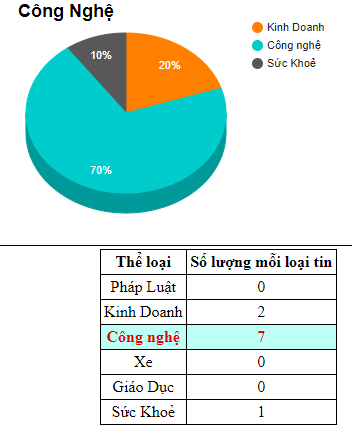
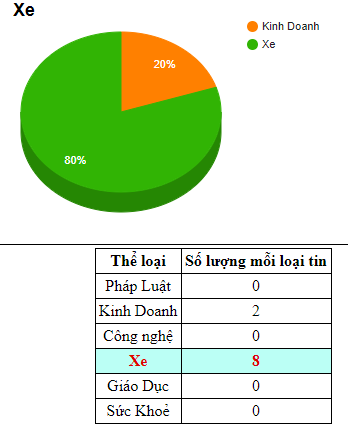
* Trang thiết lập tài khoản

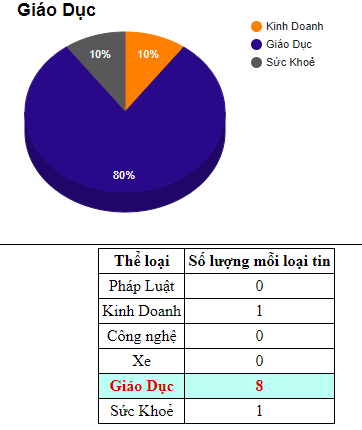
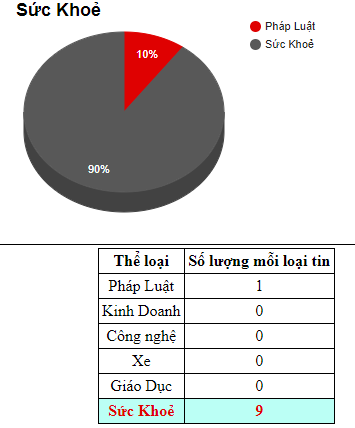


Hình 4.21 Trang thiết lập tài khoản

## **4.3 Giao diện đồ thị đánh giá độ chính xác của thuật toán:**



Hình 2.3 Giao diện đồ thị

Ở hình trên ta thấy có 6 đồ thị, đại diện cho 6 thể loại tin tức trong hệ thống. Mỗi hình thể hiện độ chính xác khi phân loại tin tức, cũng như là xem được có bao nhiêu tin tức bị phân loại nhầm và nhầm vào thể loại nào.

Màu đỏ đại diện cho tin tức được phân loại vào thể loại “*Pháp Luật*”.

Tương tự màu cam là “Kinh Doanh”.

Màu xanh ngọc là “Công Nghệ”.

Màu xanh lá là “Xe”.

Màu xanh biển là “Giáo Dục”.

Màu xám là “Sức Khoẻ”.

Cụ thể ở đồ thị số 1, có thể thấy được đồ thị này đang phân loại 10 tin tức, trong đó:

* 9 tin cho thế loại “*Pháp* *Luật*” với tỷ lệ là 90%
* 1 tin cho thể loại “*Kinh* *Doanh*” với tỷ lệ 10%.

Như vậy có thể thấy được ở đồ thị số 1, tỷ lệ chính xác khi phân loại 10 tin tức là 90% (phân loại đúng 10 tin, phân loại nhầm 1 tin). Tương tự với 5 đồ thị còn lại, ta thấy:

Đồ thị số 2: độ chính xác 90%.

Đồ thị số 3: độ chính xác 70%.

Đồ thị số 4: độ chính xác là 80%.

Đồ thị số 5: độ chính xác là 80%.

Đồ thị số 6: độ chính xác là 90%.

# CHƯƠNG 5. KẾT LUẬN VÀ KIẾN NGHỊ

## **5.1 Kết quả đạt được**

Sau thời gian thực hiện đề tài, hệ thống tuy chưa hoàn thành nhưng tôi đã đạt được một số kết quả sau:

* Hiểu rõ quy trình xây dựng hệ thống web tin tức.
* Xây dựng thành công website xem tin tức đáp ứng nhu cầu đặt ra của người dùng cũng như ban quản trị.
* Tìm hiểu và nắm khá rõ về các ngôn ngữ, thư viện, công cụ để xây dựng website như: HTML, CSS, JS, PHP, Laravel, MySQL, NodeJS, …
* Giao diện của chương trình thân thiện, dễ sử dụng.
* Hệ thống đã giúp người dùng tiết kiệm thời gian, công sức để có thể cập nhật tin tức cuộc sống xảy ra xung quanh họ, thay vì phải đi mua một tờ báo hoặc nghe người khác kể lại. Dễ dàng quản lý thông tin chi tiết người dùng, các thông tin về tin tức và loại tin, bình luận của người dùng từ đó sẽ thuận tiện cho việc đưa tin tức nhanh và chính xác nhất đến cho bạn đọc.
* Thực hiện kiểm thử phần mềm ở mức độ lập trình.
* Thực hiện lấy dữ liệu, thông tin của các tin tức ở trang tuoitre.vn và lưu vào cơ sở dữ liệu bằng NodeJS.
* Ứng dụng thuật toán phân loại Naive Bayes Classifier vào hệ thống để phân loại dữ liệu khi lấy về từ trang web khác.
* Ứng dụng google Chart để thống kê, báo cáo độ chính xác của thuật toán phân loại.
* Tiền xử lý dữ liệu khi lấy từ web khác về và khi chuẩn bị ứng dụng thuật toán.

## **5.2 Hạn chế của đề tài**

* + Vì đi thực tập ở công ty cộng với việc cần thời gian tìm hiểu về framework Laravel, NodeJS nên thời gian thực hiện đề tài bị rút ngắn.
  + Chưa thể cập nhật hết được tất cả những tin tức nóng nổi nhất hiện đang có tại một thời điểm vì dữ liệu chỉ được lấy về từ một trang tin tức khác.
  + Phần người dùng và quản trị có một số chức năng chưa hoàn thành xong.
  + Các chức năng chưa thực sự được tối ưu hóa.

## **5.3 Hướng phát triển của đề tài**

Với nỗ lực của bản thân, tôi đã cố gắng hoàn thành yêu cầu đề tài. Do mất thời gian đi thực tập ở công ty và năng lực có hạn nên website của tôi mới chỉ đi sâu vào chức năng cơ bản cho quản trị viên và người dùng. Nên tôi muốn sau đợt này sẽ:

* + Phát triển hoàn chỉnh giao diện của user và admin.
  + Phát triển các chức năng trong phần user và admin.
  + Phát triển một hệ thống tin tức: có thể lấy được hết dữ liệu của các trang tin tức hàng đầu Việt Nam hiện nay.
  + Tối ưu hoá thời gian lấy dữ liệu từ các trang web khác.
  + Phát triển hệ thống thông báo, phản hồi cho người dùng khi họ bình luận trong website thông qua email.
  + Trở thành một website cung cấp tin tức uy tín, nhanh và chính xác. Cung cấp đầy đủ những tin tức nống nổi và thời sự nhất đang xảy ra trên toàn quốc cũng như thế giới.
  + Tìm hiểu và ứng dụng thêm các thuật toán phân loại khác vào hệ thống để có cái nhìn chính xác và cụ thể hơn cho từng loại thuật toán.

## **5.4 Đề nghị ý kiến**

Trong thời gian thực hiện đề tài này, tôi thực sự không thể tránh khỏi những thiếu sót, vì vậy rất mong nhận được ý kiến đóng góp từ phía quý thầy cô và các bạn để tôi hoàn thiện hơn hệ thống xây dựng trang web tin tức.

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | T. Otwell, "Laravel," 6 2011. [Online]. Available: https://laravel.com/docs/5.4. |
| [2] | V. T. Tài, "toidicode," 2017. [Online]. Available: https://toidicode.com/cau-truc-cua-laravel-5-3-2.html. |
| [3] | W3Schools, "W3Schools," Refsnes Data, 1998. [Online]. Available: https://www.w3schools.com/nodejs/default.asp. [Accessed 09 08 2016]. |
| [4] | dougwilson, "NPM," 11 2 2020. [Online]. Available: https://github.com/mysqljs/mysql. |
| [5] | P. H. Hoàng, Tôi Đi Code Dạo, 12 12 2017. [Online]. Available: https://toidicodedao.com/2017/12/12/puppeteer-headless-chrome-api-phan-1/. |
| [6] | P. H. Hoàng, "Tôi Đi Code Dạo," 19 12 2017. [Online]. Available: https://toidicodedao.com/2017/12/19/puppeteer-headless-chrome-api-phan-2-cao-du-lieu-kenh14/. |
| [7] | C. Howard, "NPM," 15 12 2016. [Online]. Available: https://www.npmjs.com/package/remove-punctuation. |
| [8] | V.-D. Le, "GitHub," 31 08 2019. [Online]. Available: https://github.com/stopwords/vietnamese-stopwords. |
| [9] | N. B. Vũ, "GitHub," 07 06 2019. [Online]. Available: https://github.com/vunb/vntk. |
| [10] | R. A. Kershaw, "NPM," 10 05 2020. [Online]. Available: https://www.npmjs.com/package/whichx. |
| [11] | thuynt, "HocTrucTuyen123.NET," 21 07 2018. [Online]. Available: http://hoctructuyen123.net/tong-quan-ve-thuat-toan-phan-lop-naive-bayes-classification-nbc/. |
| [12] | N. T. Hop, "VIBLO," 14 09 2019. [Online]. Available: https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM. |