

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT
VIỆN KỸ THUẬT CÔNG NGHỆ**



BÁO CÁO TỐT NGHIỆP

Đề tài:

**XÂY DỰNG BẢNG TIN RÚT GỌN VỀ DỊCH BỆNH COVID-19
BẰNG KỸ THUẬT TEXT MINING**

Người hướng dẫn: Ths **DƯƠNG THỊ KIM CHI**

Người thực hiện: **HOÀNG KIM TUYẾN** – 1824801040043

Lớp: D18HT01

Niên khóa: 2018 - 2022

BÌNH DƯƠNG , NĂM 2022

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT
VIỆN KỸ THUẬT CÔNG NGHỆ**



BÁO CÁO TỐT NGHIỆP

Đề tài:

**XÂY DỰNG BẢNG TIN RÚT GỌN VỀ DỊCH BỆNH COVID-19
BẰNG KỸ THUẬT TEXT MINING**

Người hướng dẫn: Ths.**ĐƯƠNG THỊ KIM CHI**

Người thực hiện: **HOÀNG KIM TUYẾN – 1824801040043**

Lớp: **D18HT01**

BÌNH DƯƠNG , NĂM 2022

PHẦN A: GIỚI THIỆU

LỜI CẢM ƠN

Trước tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc đến Cô Ths. Dương Thị Kim Chi, người đã tận tình chỉ bảo, hướng dẫn tôi trong suốt quá trình thực hiện đề án tốt nghiệp này.

Tôi xin bày tỏ lời cảm ơn sâu sắc đến các thầy cô giảng viên, thư kí chương trình đã luôn đôn đốc nhắc nhở, hỗ trợ và giảng dạy tôi trong suốt bốn năm học qua, đã cho tôi nhiều kiến thức quý báu để tôi vững bước trên con đường học tập của mình.

Tôi xin gửi lời cảm ơn tới các bạn trong khoá D18HT- ngành Công nghệ thông tin đã đồng hành, giúp đỡ tôi trong suốt quá trình học tập tại trường.

Xin chân thành cảm ơn!

Thủ Dầu Một, ngày 03 tháng 05 năm 2022

Người thực hiện đề tài

Hoàng Kim Tuyền

QUYẾT ĐỊNH GIAO ĐỀ TÀI

UBND TỈNH BÌNH DƯƠNG
TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT

Số: 26 /QĐ - ĐHTDM

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Bình Dương, ngày 06 tháng 01 năm 2022

QUYẾT ĐỊNH

QUYẾT ĐỊNH:

Điều 1. Giao đề tài Báo cáo tốt nghiệp và giảng viên hướng dẫn cho sinh viên các Chương trình:

1. Chương trình Giáo dục Tiểu học: 62 báo cáo tốt nghiệp;
2. Chương trình Giáo dục Mầm non: 36 báo cáo tốt nghiệp;
3. Chương trình Sư phạm Ngữ văn: 17 báo cáo tốt nghiệp;
4. Chương trình Ngôn ngữ Anh: 58 báo cáo tốt nghiệp;
5. Chương trình Ngôn ngữ Trung Quốc: 18 báo cáo tốt nghiệp;
6. Chương trình Chính trị học: 03 báo cáo tốt nghiệp;
7. Chương trình Luật: 21 báo cáo tốt nghiệp;
10. Chương trình Kiến trúc: 16 báo cáo tốt nghiệp;
11. Chương trình Văn hóa học: 17 báo cáo tốt nghiệp;

12. Chương trình Quản trị Kinh doanh: 15 báo cáo tốt nghiệp;
13. Chương trình Kỹ thuật Điện – Điện tử: 51 báo cáo tốt nghiệp;
14. Chương trình Công nghệ thông tin: 49 báo cáo tốt nghiệp;
15. Chương trình Trí tuệ nhân tạo & Hệ thống thông tin: 29 báo cáo tốt nghiệp;
16. Chương trình Vật lý: 15 báo cáo tốt nghiệp.

(Danh sách kèm theo)

Điều 2. Giảng viên, sinh viên có tên trong danh sách chịu trách nhiệm thực hiện và hoàn thành đề tài Báo cáo tốt nghiệp theo quy định.

Điều 3. Giám đốc Trung tâm Đảm bảo chất lượng, Trưởng Phòng/Trưởng Khoa/ Viện/Giám đốc Chương trình và các cá nhân có tên tại Điều 1 chịu trách nhiệm thi hành Quyết định này.

Quyết định này có hiệu lực kể từ ngày ký./.

Nơi nhận:

- HT, các PHT;
- Như Điều 3;
- Lưu VT, ĐBCL.

KT. HIỆU TRƯỞNG
PHÓ HIỆU TRƯỞNG



TS. Ngô Hồng Diệp

UBND TỈNH BÌNH DƯƠNG
TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

**DANH SÁCH GIAO ĐỀ TÀI VÀ GIẢNG VIÊN HƯỚNG DẪN THỰC HIỆN BÁO CÁO TỐT NGHIỆP
CHƯƠNG TRÌNH TRÍ TUỆ NHÂN TẠO & HỆ THỐNG THÔNG TIN, ĐỢT 1, HỌC KÌ 2, NĂM HỌC: 2021 - 2022**

Kèm theo Quyết định số 26 / QĐ- ĐHTDM ngày 06 tháng 04 năm 2022 của Hiệu trưởng Trường Đại học Thủ Dầu Một

STT	Họ tên sinh viên	MSSV	Lớp	Tên đề tài	GVHD
11	Hoàng Kim Tuyền	1824801040043	D18HT01	Xây dựng bảng tin rút gọn về dịch bệnh Covid-19 bằng kỹ thuật Text mining	ThS. Dương Thị Kim Chi
12	Đặng Trọng Nghĩa	1824801040056	D18HT01	Xây dựng ứng dụng quản lý lịch học của sinh viên trường ĐHTDM	ThS. Dương Thị Kim Chi
13	Nguyễn Thị Hạnh Dung	1824801040020	D18HT01	Ứng dụng học máy xây dựng hệ thống điểm danh bằng nhận diện khuôn mặt trong kỳ thi trực tuyến trên Microsoft Team	ThS. Dương Thị Kim Chi
14	Nguyễn Đức Long	1824801040033	D18HT01	Xây dựng website tư vấn món ăn GoodHealthy	ThS. Dương Thị Kim Chi

TÓM TẮT

Trích rút từ khoá từ trang web là một bài toán hay của hệ thống bài toán trích rút từ khoá cho một văn bản. Ở mức cao hơn, nó là một bài toán con trong hệ thống trích xuất thông tin (Information Retrieval). Trong nhiều năm qua, bài toán này đã được đề cập, quan tâm nhiều ở các hội nghị quốc tế và các công ty lớn. Bài toán trích rút từ khoá từ trang web là việc trích rút từ khoá trong văn bản nội dung trang web. Đây cũng là vấn đề khá mới mẻ và được áp dụng trong rất nhiều lĩnh vực khác nhau như: Hỗ trợ tìm kiếm, hỗ trợ gợi ý người dùng, tóm tắt văn bản, v.v.

Việc đọc và tóm tắt nội dung của các văn bản trên Internet rất khó khăn và tốn nhiều thời gian cho con người, đến mức gần như không thể đạt được với nguồn nhân lực hạn chế khi kích thước của thông tin tăng lên. Kết quả là các hệ thống tự động thường được sử dụng để thực hiện nhiệm vụ này. Sự ra đời của các máy tìm kiếm đã phần nào giải quyết được vấn đề tràn ngập thông tin của các trang web. Các máy tìm kiếm chủ yếu vẫn sử dụng những từ khoá và tìm những trang có chứa từ khoá và cho ra kết quả phù hợp.

Việc trích chọn từ khoá là ứng dụng quan trọng nhất trong các engine tìm kiếm. Vì hiện nay các engine này chủ yếu vẫn tìm kiếm dựa vào từ khoá. Đó chính là một trong những động lực để phát triển bài toán trích rút từ khoá từ trang web. Nhiệm vụ bài toán đặt ra là cần tìm được một tập các từ khoá sao cho các từ khoá này phải sát với nội dung của tài liệu văn bản. Vì thế các phương pháp tóm tắt tự động được nghiên cứu và phát triển.

Trong đồ án tốt nghiệp này, người thực hiện đề tài đã nghiên cứu và thực hiện các phương pháp trích rút từ khoá và tóm tắt văn bản trang web mang nội dung về dịch bệnh Covid-19 sử dụng kỹ thuật Text mining, nhằm xây dựng một bản tin Covid-19 với nội dung ngắn gọn, giúp người đọc tiết kiệm thời gian và nắm được sơ bộ nội dung chính của bài báo, giúp cho việc tìm hiểu và tổng hợp thông tin được dễ dàng hơn.

SUMMARY

Extracting keywords from web pages is a good problem in the system of keyword extraction problems for a document. At a higher level, it is a subproblem in the Information Retrieval system. For many years, this problem has been mentioned and interested in many international conferences and big companies. The problem of extracting keywords from the website is the extraction of keywords in the web content text. This is also a fairly new problem and is applied in many different fields such as: Search support, user suggestion support, text summaries, etc.

Reading and summarizing the content of texts on the Internet is difficult and time consuming for humans, to the point of becoming almost impossible with limited human resources as the size of the information increases. As a result, automated systems are often used to perform this task. The advent of search engines has partly solved the problem of information flooding of the web. Search engines still mostly use keywords and find pages that contain keywords and return relevant results.

Keyword extraction is the most important application in search engines. Because nowadays these engines are still searching based on keywords. That is one of the motivations to develop the problem of extracting keywords from the website. The problem is to find a set of keywords so that these keywords must be close to the content of the text document. Therefore, automatic summarization methods are researched and developed.

In this graduation project, the researcher has researched and implemented methods of keyword extraction and text summarization of websites with content about the Covid-19 epidemic using Text mining technique, in order to build build a Covid-19 newsletter with concise content, helping readers save time and preliminary grasp the main content of the article, making it easier to learn and synthesize information.

MỤC LỤC

PHẦN A: GIỚI THIỆU	i
LỜI CẢM ƠN	i
QUYẾT ĐỊNH GIAO ĐỀ TÀI	i
TÓM TẮT	v
SUMMARY	vi
MỤC LỤC.....	vii
DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT	ix
DANH MỤC CÁC BẢNG.....	x
DANH MỤC CÁC HÌNH VẼ	xi
PHẦN B: NỘI DUNG	1
CHƯƠNG 1: TỔNG QUAN BÀI TOÁN.....	1
1.1 Lý do chọn đề tài.....	1
1.2 Mục đích nghiên cứu.....	2
1.3 Đối tượng nghiên cứu và phạm vi nghiên cứu.....	2
1.4 Bố cục của đồ án	2
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	3
2.1 Ngôn ngữ lập trình Python.....	3
2.2 Text Mining.....	4
2.2.1 Thu thập dữ liệu web sử dụng thư viện BeautifulSoup(Python)	4
2.2.2 Xử lý ngôn ngữ tự nhiên(NLP).....	5
2.2.3 Thuật toán tóm tắt văn bản	6
2.2.3.1. Một số khái niệm cơ bản.....	6
2.2.3.2 Phân loại bài toán tóm tắt	7
2.2.3.3 Tóm tắt văn bản sử dụng phương pháp Textrank	10
2.3 Framework Django	14

2.3.1 Ưu điểm của Django	15
2.3.2 Nhược điểm của Django	15
2.3.3 MVT Pattern của Django	16
2.3.4 Các thành phần cơ bản của ứng dụng Django	17
2.3.5 Lý do chọn Django cho đề tài.....	18
CHƯƠNG 3: MÔ HÌNH ĐỀ XUẤT	20
3.1 Thu thập dữ liệu web sử dụng BeautifulSoup.....	20
3.2 Tiền xử lý và làm sạch dữ liệu(NLP).....	20
3.3 Mô hình tổng quát.....	22
3.3.1 Mô hình tổng quát thu thập dữ liệu.....	22
3.3.2 Tổng quát quá trình xử lý ngôn ngữ tự nhiên	24
3.3.3 Use Case của bảng tin thu gọn.....	25
CHƯƠNG 4: THỰC NGHIỆM	27
4.1 Các công nghệ và thư viện sử dụng	27
4.2 Mô tả dữ liệu thu thập	27
4.3 Xây dựng bảng tin thu gọn về dịch bệnh Covid-19(sử dụng Django Framework).....	29
4.3.1 Khởi tạo project Django.....	29
4.3.2 Thu thập dữ liệu covid-19.....	30
4.3.3 Xây dựng giao diện ứng dụng.....	34
4.3.4 Giao diện và chức năng trang quản trị	39
4.3.5 Các chức năng của ứng dụng cho người dùng.....	44
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	46
5.1 Kết luận	46
5.2 Hướng phát triển	46
PHẦN C: PHỤ LỤC VÀ TÀI LIỆU THAM KHẢO	47

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC CHỮ VIẾT TẮT

Ths (Thạc sĩ)

NLP (Natural Language Processing)

IR (Information Extraction)

MVT (Model-View-Template)

MVC (Model-View-Controller)

DANH MỤC CÁC BẢNG

Bảng 1: Mô tả bảng dữ liệu covid_links_news.....	28
Bảng 2: Mô tả bảng dữ liệu covid_news.....	28

DANH MỤC CÁC HÌNH VẼ

Hình 2.1 Framework Django	14
Hình 2.2 Mô hình MVT của Django.....	16
Hình 3.1 Mô tả tổng quát ứng dụng	22
Hình 3.2 Use case của người dùng.....	25
Hình 3.3 Use case của Admin	26
Hình 4.1 Các thuộc tính trong bảng dữ liệu	27
Hình 4.2 Cấu trúc project	29
Hình 4.3 Cấu trúc thư mục crawl_data	30
Hình 4.4 phân tích cấu trúc HTML trang chứa link bài báo	31
Hình 4.5 Phân tích cấu trúc HTML của bài báo	32
Hình 4.6 Dữ liệu sau khi thu thập	33
Hình 4.7 Dữ liệu bài báo sau khi thu thập và xử lí	34
Hình 4.8 Trang chủ của ứng dụng.....	34
Hình 4.9 Trang chủ tìm kiếm của ứng dụng	35
Hình 4.10 Bảng tin Covid-19	36
Hình 4.11 Giao diện bài viết thu gọn nội dung về Covid-19	37
Hình 4.12 Giao diện phần dưới bài viết.....	37
Hình 4.13 Giao diện footer của bảng tin.....	38
Hình 4.14 Giao diện đăng nhập trang admin	39
Hình 4.15 Giao diện đăng nhập trang admin	39
Hình 4.16 Giao diện trang quản trị	40
Hình 4.17 Giao diện quản trị link bài viết.....	41
Hình 4.18 Giao diện thêm/xóa/sửa link bài viết	41
Hình 4.19 Giao diện quản trị các bài viết	42
Hình 4.20 Giao diện thêm/xóa/sửa một bài viết	43
Hình 4.21 Giao diện thêm/xóa/sửa một bài viết	43

Hình 4.22 Giao diện chức năng tìm kiếm bằng từ khóa	44
Hình 4.23 Giao diện chức năng lọc bài viết theo mặt báo.....	45
Hình 4.24 Giao diện chức năng chuyển hướng tới link gốc bài báo.....	45

PHẦN B: NỘI DUNG

CHƯƠNG 1: TỔNG QUAN BÀI TOÁN

1.1 Lý do chọn đề tài

Trích rút từ khoá tự động từ trang web là một trong những bài toán khó thuộc hệ bài toán tóm tắt văn bản. Hiện nay trên thế giới, có rất nhiều nhà khoa học và các công ty tỏ ra rất quan tâm đến bài toán trích rút từ khoá tự động. Tại các hội nghị nổi tiếng như DUC 2001 – 2007, TAC 2008 – 2011, ACL 2001 – 2015, trích rút từ khoá tự động đã được đề cập đến nhiều trong các bài báo. Ngoài ra, có nhiều hệ thống tóm tắt văn bản độc lập hoặc tích hợp được phát triển như: MEAD, LexRank, chức năng tự động tóm tắt của Microsoft Word.

Các bài báo về dịch bệnh Covid cũng không nằm ngoài số đó. Một tình trạng mà các trang báo nào cũng từng gặp phải đó là việc đăng tải thông tin nội dung quá dài, khiến cho người đọc có thể xảy ra trạng thái lười đọc hoặc không đọc vì nội dung của bài báo quá dài và không tập trung vào nội dung chính, gây khó khăn trong việc tìm hiểu và tổng hợp thông tin.

Với thực tế nêu trên, người thực hiện đề tài đã đề xuất một giải pháp giải quyết bài toán trích xuất từ khóa trang web qua đề tài “**Xây dựng bảng tin rút gọn về dịch bệnh Covid-19 bằng kỹ thuật Text mining**” nhằm tạo ra một hệ thống hỗ trợ người đọc trích xuất nội dung thông của bài báo covid-19 một cách ngắn nhất mà vẫn giữ được nội dung chính của bài báo, hệ thống sẽ sử dụng các kỹ thuật trong Text Mining để trích xuất và tóm tắt nội dung, giúp người đọc hiểu được sơ bộ nội dung chính của bài báo mà họ sắp đọc.

1.2 Mục đích nghiên cứu

Đề tài “**Xây dựng bảng tin rút gọn về dịch bệnh Covid-19 bằng kỹ thuật Text mining**”, là đề tài hướng tới xây dựng bảng tin covid-19 ngắn gọn, cung cấp thông tin bài báo dưới dạng thu nhỏ, rút gọn nhưng vẫn giữ được nội dung chính của bài báo.

1.3 Đối tượng nghiên cứu và phạm vi nghiên cứu

Với mục đích xây dựng bảng tin rút gọn về covid-19, đề tài xác định các đối tượng và phạm vi nghiên cứu như sau:

- Đối tượng nghiên cứu là: các phương pháp và thuật toán trong Text Mining, thông tin, hình ảnh, số liệu các bài báo trên các báo online
- Phạm vi nghiên cứu: các bài báo đưa thông tin về dịch bệnh covid-19

1.4 Bố cục của đồ án

Nội dung của đồ án được chia thành bố cục như sau:

Chương 1: Tổng quan bài toán

Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan

Chương 3: Mô hình tổng quát

Chương 4: Thực nghiệm

Chương 5: Kết luận và hướng phát triển

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Ngôn ngữ lập trình Python

Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level) ngữ nghĩa động. Python là ngôn ngữ lập trình hướng đối tượng thông dụng dùng để viết các tiện ích hệ thống. Nó cũng được sử dụng như ngôn ngữ kết dính đóng vai trò tích hợp C và C++.

Được tạo ra bởi Guido van Rossum tại Amsterdam năm 1990, Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý.

Python là ngôn ngữ có hình thức khá đơn giản và rõ ràng, do đó tạo nên sự dễ dàng tiếp cận cho những lập trình viên mới bắt đầu.

Ban đầu, Python được phát triển để chạy trên nền Unix, nhưng rồi theo thời gian, nó đã được mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Windows, Linux và các hệ điều hành khác thuộc họ Unix.

Một số tính năng nổi bật của python như:

- Miễn phí, mã nguồn mở
- Ngôn ngữ lập trình đơn giản, dễ đọc
- Khả năng di chuyển
- Khả năng mở rộng và có thể nhúng
- Ngôn ngữ thông dịch cấp cao
- Hướng đối tượng
- Thư viện tiêu chuẩn lớn
- Khoa học, phân tích số liệu

2.2 Text Mining

Khai thác văn bản (text mining hoặc text data mining) là một quá trình xử lý và trích xuất thông tin nằm trong văn bản, quá trình này là một phần của việc phân tích văn bản trong khai thác dữ liệu. Thông tin được thể hiện dưới dạng các mẫu, xu hướng, thứ tự sắp xếp được trích xuất thông qua các luật hoặc thông qua quá trình học dựa trên các mẫu thống kê.

Khai thác văn bản bao gồm các bước cơ bản như: tiền xử lý, học mô hình, phán đoán, tổng hợp phân tích và trình bày kết quả. Tiền xử lý có thể gồm việc phân tách đoạn văn bản thành các đoạn nhỏ hơn, làm giàu văn bản bằng các tri thức bên ngoài, hoặc loại bỏ những thông tin nhiễu trong văn bản. Quá trình học là quá trình tìm ra các mẫu trong một tập các văn bản đã được tiền xử lý hoặc chưa qua tiền xử lý, kết quả quá trình học là một mô hình biểu diễn các mẫu được tìm thấy.

Quá trình phán đoán là quá trình áp dụng mô hình vừa học được trên các văn bản mới, văn bản mới sẽ được gán nhãn thêm thông tin. Cuối cùng là quá trình tổng hợp và trình bày kết quả. Khai phá văn chia thành các vấn đề nhỏ hơn bao gồm phân loại tài liệu (text categorization, text classification), gom cụm văn bản (text clustering), trích xuất thực thể (concept/entity extraction), phân tích tình cảm (sentiment analysis), tóm tắt tài liệu (document summarization), và trích xuất quan hệ giữa các thực thể (entity relation modeling).

2.2.1 Thu thập dữ liệu web sử dụng thư viện BeautifulSoup(Python)

Beautiful Soup là một gói Python để phân tích cú pháp các tài liệu HTML và XML. Nó tạo ra một cây phân tích cú pháp cho các trang được phân tích cú pháp có thể được sử dụng để trích xuất dữ liệu từ HTML, rất hữu ích cho việc cào dữ liệu trên website.

2.2.2 Xử lý ngôn ngữ tự nhiên(NLP)

Xử lý ngôn ngữ tự nhiên là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người, dưới dạng tiếng nói (speech) hoặc văn bản (text). Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

Xử lý ngôn ngữ tự nhiên có thể được chia ra thành hai nhánh lớn, không hoàn toàn độc lập, bao gồm xử lý tiếng nói (speech processing) và xử lý văn bản (text processing). Xử lý tiếng nói tập trung nghiên cứu, phát triển các thuật toán, chương trình máy tính xử lý ngôn ngữ của con người ở dạng tiếng nói (dữ liệu âm thanh). Các ứng dụng quan trọng của xử lý tiếng nói bao gồm nhận dạng tiếng nói và tổng hợp tiếng nói. Nếu như nhận dạng tiếng nói là chuyển ngôn ngữ từ dạng tiếng nói sang dạng văn bản thì ngược lại, tổng hợp tiếng nói chuyển ngôn ngữ từ dạng văn bản thành tiếng nói.

Xử lý văn bản tập trung vào phân tích dữ liệu văn bản. Các ứng dụng quan trọng của xử lý văn bản bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động, hay kiểm lỗi chính tả tự động. Xử lý văn bản đôi khi được chia tiếp thành hai nhánh nhỏ hơn bao gồm hiểu văn bản và sinh văn bản. Nếu như hiểu liên quan tới các bài toán phân tích văn bản thì sinh liên quan tới nhiệm vụ tạo ra văn bản mới như trong các ứng dụng về dịch máy hoặc tóm tắt văn bản tự động.

Các bài toán và ứng dụng của NLP:

- Phân loại tài liệu (Text classification) là quá trình gán các danh mục được xác định trước cho tài liệu có sẵn dựa trên nội dung của nó. Điều này có thể thực hiện thủ công hay sử dụng các thuật toán khác nhau. Cho đến nay, phân loại tài liệu là ứng dụng phổ biến nhất của NLP, được sử dụng để phát triển các công cụ khác nhau như trình phát hiện thư rác (Spam),...

- Trích xuất thông tin (Information extraction – IE) là quá trình tự động trích xuất thông tin có liên quan từ các tài liệu dạng văn bản không có cấu trúc và / hoặc bán cấu trúc. Ví dụ về các loại tài liệu này bao gồm các sự kiện lịch từ email hoặc danh sách những tài khoản được gắn thẻ trong một bài đăng trên mạng xã hội,...
- Truy xuất thông tin (Information retrieval – IR) là nhiệm vụ hỗ trợ tìm kiếm các tài liệu được người dùng yêu cầu truy vấn từ một cơ sở dữ liệu lớn có liên quan, chẳng hạn như thanh tìm kiếm của Google.
- Tự động tóm tắt văn bản là quá trình rút ngắn một tập hợp dữ liệu về mặt tính toán, để tạo ra một tập hợp con đại diện cho thông tin quan trọng nhất hoặc có liên quan trong nội dung gốc.
- Khai phá dữ liệu (data mining) và phát hiện tri thức: Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình học tập, khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển.
- Kiểm lỗi chính tả tự động là việc sử dụng máy tính để tự động phát hiện các lỗi chính tả trong văn bản (lỗi từ vựng, lỗi ngữ pháp, lỗi ngữ nghĩa) và đưa ra gợi ý cách chỉnh sửa lỗi.

2.2.3 Thuật toán tóm tắt văn bản

2.2.3.1. Một số khái niệm cơ bản

- Tỷ lệ nén (Compression Rate): là độ đo giữa thông tin văn bản tóm tắt và văn bản gốc được tính bằng công thức:

$$CompressionRate = \frac{SummaryLength}{SourceLength}$$

Trong đó:

- + SummaryLength: Độ dài văn bản tóm tắt
- + SourceLength: Độ dài văn bản gốc
- Độ liên quan (Relevance): là độ đo cho mức độ quan trọng của thông tin

tin mà văn bản tóm tắt có được so với văn bản gốc.

- Sự mạch lạc (Coherence): là thước đo cho sự mạch lạc, tuân theo thể thống nhất của văn bản, không có sự trùng lặp các thành phần.

2.2.3.2 Phân loại bài toán tóm tắt

Có nhiều cách phân loại tóm tắt văn bản khác nhau, phụ thuộc vào các yếu tố cơ bản sau:

- + Định dạng văn bản, nội dung đầu vào
- + Định dạng, nội dung đầu ra
- + Mục đích tóm tắt

- Phân loại tóm tắt dựa trên định dạng, nội dung đầu vào:

- Kiểu văn bản (bài báo, bản tin, thư, báo cáo ...). Với cách phân loại này, tóm tắt văn bản là bài báo sẽ khác với tóm tắt thư, tóm tắt báo cáo khoa học do những đặc trưng văn bản quy định.
- Định dạng văn bản: dựa vào từng định dạng văn bản khác nhau, tóm tắt cũng chia ra thành các loại khác nhau như: tóm tắt văn bản không theo cấu trúc nhất định và tóm tắt văn bản có cấu trúc. Đối văn bản có cấu trúc, tóm tắt văn bản thường sử dụng một mô hình học dựa vào mẫu cấu trúc đã xây dựng từ trước để tiến hành tóm tắt
- Số lượng dữ liệu đầu vào: Tóm tắt đơn văn bản khi đầu vào chỉ là một văn bản đơn, trong khi đó đầu vào của tóm tắt đa văn bản là một tập các tài liệu có liên quan đến nhau như: các tin tức có liên quan đến cùng một sự kiện, các trang web cùng chủ đề hoặc là cụm dữ liệu được trả về từ quá trình phân cụm.
- Miền dữ liệu: tùy theo miền của dữ liệu về cụ thể về một lĩnh vực nào đó, ví dụ như: y tế, giáo dục... hay miền dữ liệu tổng quát, có thể chia tóm tắt ra thành từng loại tương ứng.

- Tóm tắt trên cơ sở mục đích thực chất là làm rõ cách tóm tắt, mục đích tóm tắt là gì, tóm tắt phục vụ đối tượng nào
 - + Nếu phụ thuộc vào đối tượng đọc tóm tắt thì tóm tắt cho chuyên gia khác cách tóm tắt cho các đối tượng đọc thông thường.
 - + Tóm tắt sử dụng trong tìm kiếm thông tin sẽ khác với tóm tắt phục vụ cho việc sắp xếp.
 - + Dựa trên mục đích tóm tắt, còn có thể chia ra thành tóm tắt chỉ thị và tóm tắt thông tin. Tóm tắt chỉ thị chỉ ra loại của thông tin, ví dụ như là loại văn bản chỉ thị “tuyệt mật”. Còn tóm tắt thông tin chỉ ra nội dung của thông tin
 - + Tóm tắt trên cơ sở truy vấn (Query-based) hay tóm tắt chung. Tóm tắt chung có mục đích chính là tìm ra đoạn tóm tắt cho toàn bộ văn bản mà nội dung của đoạn văn bản sẽ bao quát toàn bộ nội dung của văn bản đó. Tóm tắt trên cơ sở truy vấn thì nội dung của văn bản tóm tắt sẽ dựa trên truy vấn của người dùng hay chương trình đưa vào, loại tóm tắt này thường được sử dụng trong quá trình tóm tắt các kết quả trả về từ máy tìm kiếm.
- Tóm tắt trên cơ sở đầu ra cũng có nhiều cách phân loại:
 - Dựa vào ngôn ngữ: Tóm tắt cũng có thể phân loại dựa vào khả năng tóm tắt các loại ngôn ngữ:
 - + Tóm tắt đơn ngôn ngữ (Monolingual): hệ thống có thể tóm tắt chỉ một loại ngôn ngữ nhất định như: tiếng Việt hay tiếng Anh...
 - + Tóm tắt đa ngôn ngữ (Multilingual): hệ thống có khả năng tóm tắt nhiều loại văn bản của các ngôn ngữ khác nhau, tuy nhiên tương ứng với văn bản đầu vào là ngôn ngữ gì thì văn bản đầu ra cũng là ngôn ngữ tương ứng.
 - + Tóm tắt xuyên ngôn ngữ (Crosslingual): hệ thống có khả năng đưa ra các văn bản đầu ra có ngôn ngữ khác với ngôn ngữ của văn bản đầu vào.
 - Dựa vào định dạng đầu ra của kết quả tóm tắt: như bảng, đoạn, từ khóa

Ngoài hai cách phân loại trên, phân loại tóm tắt trên cơ sở đầu ra còn có một cách phân loại được sử dụng phổ biến là: tóm tắt theo kiểu **trích xuất** – “**extraction**” và tóm tắt theo kiểu **tóm lược ý** – “**abstraction**”.

+ Phương pháp tóm tắt trích chọn là công việc chọn ra một tập con những từ đã có, những lời nói hoặc những câu của văn bản gốc để đưa vào khuôn mẫu tóm tắt.

+ Tóm tắt theo tóm lược: là tóm tắt có kết quả đầu ra là một tóm tắt không giữ nguyên lại các thành phần của văn bản đầu vào mà dựa vào thông tin quan trọng để viết lại một văn bản tóm tắt mới.

Tóm tắt văn bản là quá trình trích rút những thông tin quan trọng nhất từ một văn bản để tạo ra phiên bản ngắn gọn, xúc tích mang đầy đủ lượng thông tin của văn bản gốc kèm theo đó là tính đúng đắn về ngữ pháp và chính tả.

Bản tóm tắt phải giữ được những thông tin quan trọng của toàn bộ văn bản chính. Bên cạnh đó, bản tóm tắt cần phải có bố cục chặt chẽ có tính đến các thông số như độ dài câu, phong cách viết và cú pháp văn bản.

Phụ thuộc vào số lượng các văn bản, kỹ thuật tóm tắt có thể chia làm hai lớp: đơn văn bản và đa văn bản.

+ Tóm tắt đơn văn bản chỉ đơn giản là rút gọn một văn bản thành một sự trình bày ngắn gọn. Trong khi đó tóm tắt đa văn bản phải rút gọn một tập các văn bản thành một sự tóm tắt.

+ Tóm tắt đa văn bản có thể xem như một sự mở rộng của tóm tắt đơn văn bản và thường dùng với thông tin chứa trong các cụm văn bản, để người dùng có thể hiểu được cụm văn bản đó. Tóm tắt đa văn bản phức tạp hơn tóm tắt đơn văn bản vì phải làm việc trên số lượng văn bản nhiều hơn.

Việc đánh giá kết quả văn bản tóm tắt là việc làm khó khăn. Việc đánh giá tự động nhằm mục đích là tìm ra được một độ đo đánh giá văn bản tóm tắt giống với đánh giá của con người nhất. Cách đánh giá tốt nhất là sử dụng ý kiến đánh giá của các chuyên gia ngôn ngữ. Nhưng đây là một phương pháp tốn kém. Vì vậy, ngoài các phương pháp đánh giá

thủ công, vấn đề đánh giá tự động kết quả tóm tắt cũng nhận được nhiều sự chú ý của các nhà nghiên cứu. Tùy vào từng loại văn bản mà mỗi cách đánh giá kết quả tóm tắt lại khác nhau. Đây là một vấn đề khá khó khăn và bất cập trong việc tìm ra độ đo thể hiện sự giống nhau về nội dung giữa văn bản tóm tắt và văn bản gốc.

2.2.3.3 Tóm tắt văn bản sử dụng phương pháp Textrank

TextRank là một kỹ thuật tóm tắt văn bản theo phương pháp extractive và trong học máy thì là học không giám sát (Unsupervised Learning). TextRank không dựa trên bất kỳ dữ liệu đào tạo nào trước đó và có thể hoạt động với bất kỳ đoạn văn bản tùy ý nào. Textrank là một thuật toán xếp hạng dựa trên biểu đồ như thuật toán PageRank của Google đã được triển khai thành công trong phân tích trích dẫn. Nó sử dụng thứ hạng văn bản thường xuyên để trích xuất từ khóa, tóm tắt văn bản tự động và xếp hạng cụm từ.

Ý tưởng của thuật toán này dựa trên hai yếu tố: bỏ phiếu và đề cử. "Khi đỉnh đầu tiên liên kết với đỉnh thứ hai, ví dụ như thông qua mối quan hệ kết nối hoặc cạnh biểu đồ. Mỗi một liên kết đến đỉnh đang xét thì nó được 1 phiếu bầu.

Như vậy, càng nhiều phiếu bầu thì đỉnh đó càng quan trọng. Từ cách xác định trên thì trọng số của một đỉnh chính là số phiếu bầu cho đỉnh đó.

Trọng số của mỗi câu u (đỉnh) được tính như sau:

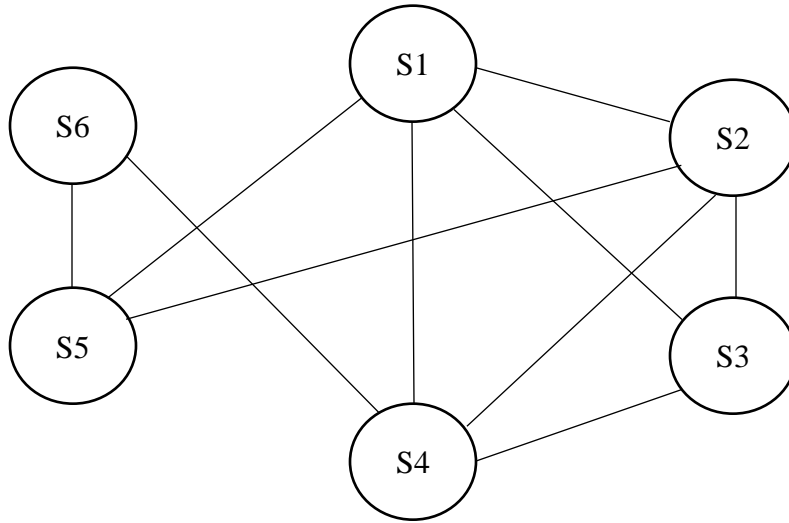
$$PageRank(u) = \frac{(1-d)}{N} + d \sum_{v \in adj[u]} \frac{PageRank(v)}{deg(v)} \quad (2.1)$$

Trong đó d là nhân tố giảm, có giá trị từ 0 đến 1. Nó là xác suất mà một đỉnh có liên kết đến một đỉnh bất kỳ trong đồ thị. Đối với các trang web thì d là xác suất người dùng nhân vào một liên kết bất kỳ và xác suất để người dùng vào một trang web hoàn toàn mới là $1 - d$.

Theo PageRank thì $d = 0.85$. Đây cũng là xác suất sẽ được sử dụng trong TextRank.

Lần đầu chạy, trọng số sẽ được gán cho các đỉnh là bằng 1.

Ta có đồ thị $G = (V, E)$ là đồ thị vô hướng. Trong đó:



V : là tập các đỉnh $\{S1, S2, S3, S4, S5, S6\}$;

E : là tập các cạnh của đồ thị

Ta có:

$$\frac{(1 - d)}{N} = \frac{1 - 0.85}{6} = 0,025$$

Trọng số mỗi đỉnh $S1 - S6$ được tính như sau:

$S1$ nối với các đỉnh $S2, S3, S4, S5$ nên trọng số sẽ bằng:

$$\text{PageRank}(S1) = 0,025 + 0,85 (1/4 + 1/3 + 1/4 + 1/3) = 1,017$$

Tương tự ta có trọng số các đỉnh khác như sau:

$$\text{PageRank}(S2) = 0,025 + 0,85 (1/4 + 1/4 + 1/3 + 1/3) = 1,017$$

$$\text{PageRank}(S3) = 0,025 + 0,85 (1/4 + 1/4 + 1/4) = 0,6625$$

$$\text{PageRank}(S4) = 0,025 + 0,85 (1/4 + 1/4 + 1/3 + 1/2) = 1,158$$

$$\text{PageRank}(S5) = 0,025 + 0,85 (1/4 + 1/4 + 1/2) = 0,875$$

- $\text{PageRank}(S5) = 0,025 + 0,85 (1/3 + 1/4 + 1/2) = 0,9458$

Lần chạy đầu tiên, trọng số sẽ được gán cho các đỉnh bằng 1.

Ban đầu gán cho tất cả các đỉnh trong đồ thị các giá trị khởi tạo và tính toán lặp lại cho đến khi kết quả hội tụ lại đạt ngưỡng xác định. Sau quá trình tính toán thì trọng số của mỗi đỉnh chính là mức độ quan trọng của đỉnh đó trong toàn đồ thị.

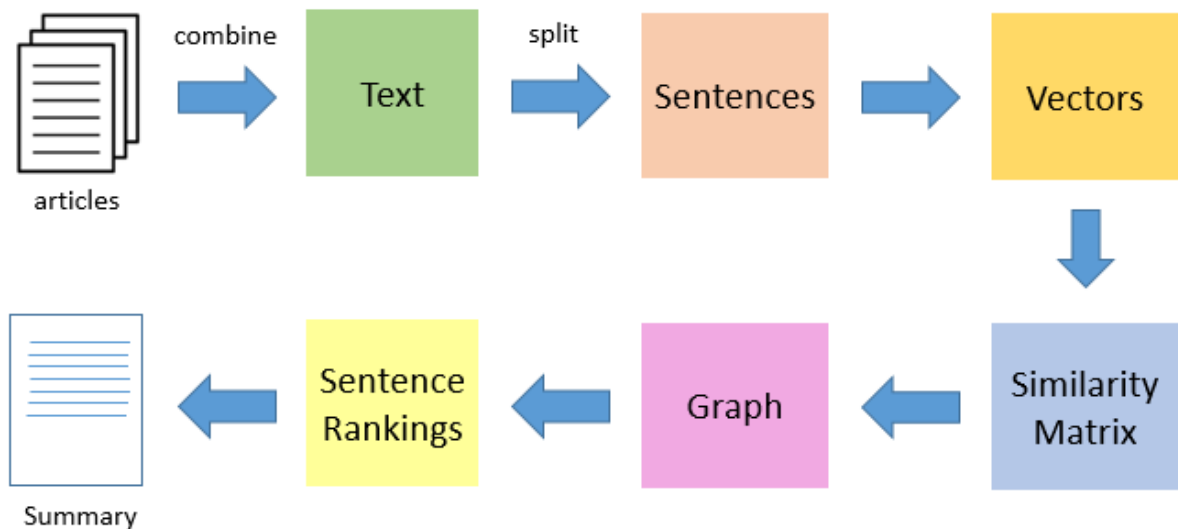
- Sử dụng TextRank trong rút trích từ khóa, tóm tắt văn bản:

Mục đích của việc trích xuất từ khoá tự động là tìm ra các cụm từ mô tả văn bản tốt nhất.

Rút trích từ khóa là chuỗi của một hoặc nhiều từ vựng được rút ra từ văn bản.

Quan hệ nằm giữa 2 đơn vị từ vựng hữu ích cho việc đánh giá thì đều được thêm vào là cạnh của đồ thị.

Tổng quan các bước trích rút, tóm tắt văn bản áp dụng TextRank như sau:



- + Bước đầu tiên sẽ là nối tất cả văn bản có trong các bài báo tạo thành 1 đoạn văn bản
- + Sau đó chia văn bản thành các câu riêng lẻ
- + Tiếp theo là biểu diễn vector cho văn bản
- + Tạo ma trận về sự tương đồng giữa các câu

- + Chuyển đổi ma trận thành một đồ thị với các câu là đỉnh và điểm là các cạnh. Từ đó tính toán thứ hạng cho các đỉnh, rồi sắp xếp theo thứ tự.
- + Cuối cùng là chọn ra các đỉnh (là các câu) được xếp hạng cao nhất để tạo thành văn bản tóm tắt.

2.3 Framework Django



Hình 2.1 Framework Django

Django là một trong số những web framework bậc cao miễn phí, là mã nguồn mở được tạo ra bởi ngôn ngữ Python dựa trên mô hình mô hình MTV (gồm Model-Template-Views). Hiện framework này được phát triển, quản lý bởi Django Software Foundation.

Django ra đời với mục tiêu hỗ trợ thiết kế các website phức tạp dựa trên những CSDL có sẵn. Nó hoạt động dựa theo nguyên lý ‘cắm’ các thành phần và tái sử dụng để tạo nên các website với ít code, ít khớp nối, có khả năng phát triển và không bị trùng lặp.

2.3.1 Ưu điểm của Django

Lợi thế hàng đầu của Django là khả năng thiết kế, tạo lập website và các ứng dụng nhanh chóng đến bất ngờ. Ngoài ra, dưới đây Bizfly liệt kê những điểm cộng khiến framework này trở nên nổi bật và được lập trình viên sử dụng rộng rãi.

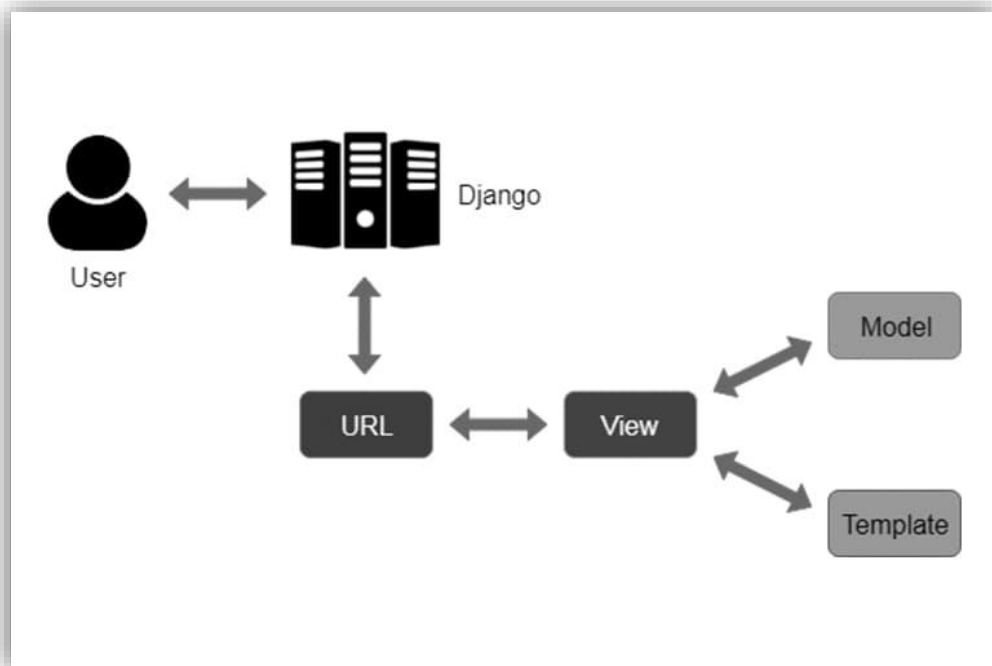
- Bảo mật tốt: Làm việc với Django, các lập trình viên gần như không có không gian để phạm bất cứ sai lầm về an ninh nào. Nó giúp các developer tránh được tất cả các lỗi thường gặp như nhấp chuột, kịch bản chéo trang, SQL tiêm, giả mạo yêu cầu,... Nhờ đó, sản phẩm được tạo bởi framework này có khả năng bảo mật cực tốt.
- Mở rộng thỏa thích: Django có sẵn tính năng mở rộng nhằm hỗ trợ các lập trình viên quản lý lưu lượng người truy cập, thích hợp với các trang có traffic lớn.
- Dễ sử dụng: Django được tạo ra bởi ngôn ngữ lập trình **Python** và mô hình MVC nên rất dễ ứng dụng trong các dự án. Đa ngôn ngữ và được hỗ trợ Multi-Site
- Dễ học: Có nhiều tài liệu hỗ trợ việc học Django, bao gồm cả tài liệu miễn phí trên mạng và sách in. Cộng đồng sử dụng Django hiện nay đang phát triển mạnh mẽ, newbie có thể tìm kiếm sự giúp đỡ trên các trang facebook, diễn đàn, blog,...

2.3.2 Nhược điểm của Django

Bên cạnh những ưu điểm kể trên, Django cũng tồn tại một số nhược điểm mà cần cân nhắc trước khi sử dụng chúng. Cụ thể sau đây:

- Có thể xảy ra một vài vấn đề khi phát triển các ứng dụng/website quy mô nhỏ
- Định tuyến tương đối khó
- Không đưa ra các cảnh báo khi xuất hiện lỗi trong mẫu.

2.3.3 MVT Pattern của Django



Hình 2.2 Mô hình MVT của Django

Django sử dụng mô hình MVT (Model-View-Template) thay vì sử dụng mô hình MVC (Model-View-Controller).

Mô hình MVT được sử dụng trong khi tạo một ứng dụng với Tương tác người dùng.

Mô hình này thì bao gồm code HTML với Django Template Language (DTL).

Controller là mã được viết để kiểm soát sự tương tác giữa Model và View và Django để đảm bảo chăm sóc nó.

Bất cứ khi nào người dùng gửi request, nó xử lý request của người dùng đó bằng Model, View và Template.

Nó hoạt động như một Controller để kiểm tra xem nó có khả dụng hay không bằng cách ánh xạ URL và nếu URL ánh xạ thành công thì View sẽ bắt đầu tương tác với Model và gửi lại Template cho người dùng dưới dạng response.

2.3.4 Các thành phần cơ bản của ứng dụng Django

Dự án Django tạo ra một tập hợp các cài đặt bao gồm cấu hình cơ sở dữ liệu và các tùy chọn cụ thể cũng như các cài đặt cụ thể của ứng dụng mặc định trong dự án. Cấu trúc thư mục của django project như sau:

- *mysite/*
- *manage.py*
- *mysite/*
- *__init__.py*
- *settings.py*
- *urls.py*
- *wsgi.py*

Trong đó

- *manager.py* Cho phép tương tác với dự án Django theo các cách khác nhau
- *__init__.py*: Nói với trình thông dịch python là thư mục nên được coi là một python package. Tập tin này chủ yếu là trống.
- *settings.py*: Tập tin cấu hình
- *urls.py*: Bao gồm tất cả khai báo URL cho dự án Django và mục lục của trang web Django.

- wsgi.py: Đây là lối vào cho các máy chủ web tương thích WSGI để phục vụ các dự án và deploy với WSGI.

2.3.5 Lí do chọn Django cho đề tài

Django được xây dựng để giúp phát triển nhanh chóng với thiết kế sạch sẽ và thiết thực. Khả năng dễ đọc của Python, đơn giản, đầy đủ của Django cho phép tập trung vào các vấn đề phức tạp, logic nghiệp vụ hơn là mất nhiều thời gian cho các rắc rối đã được người khác giải quyết. Và Django có một mô hình xác thực người dùng rất tốt với khả năng cấu hình người dùng. Điều này đã làm cho nó trở thành lựa chọn hàng đầu khi trang web, ứng dụng cần ưu tiên về bảo mật.

Nó sử dụng một loạt các thành phần Python là các thực thể riêng biệt không phụ thuộc vào nhau.

Nhận thấy lợi ích và ưu điểm của Django, người thực hiện đề tài quyết định sử dụng framework này để xây dựng website bảng tin thu gọn về dịch bệnh covid-19 của mình.

CHƯƠNG 3:MÔ HÌNH ĐỀ XUẤT

3.1 Thu thập dữ liệu web sử dụng BeautifulSoup

BeautifulSoup là một thư viện phổ biến của python giúp thu thập dữ liệu trang web một cách tự động, đơn giản mà vô cùng mạnh mẽ.

Các bước đơn giản để bắt đầu lấy một tiêu đề của bài báo bằng BeautifulSoup như sau

```
from bs4 import BeautifulSoup
url = "https://baomoi.net"
response = requests.get(url)
soup = BeautifulSoup(response.content, "html.parser")
title = soup.find('h1', class_="title")
```

Tương tự với các nội dung khác, tùy vào trang web và cấu trúc HTML của nó, để thu thập dữ liệu cụ thể của 1 trang web, cần tập trung thời gian vào nghiên cứu cấu trúc HTML của nó, các thẻ chứa dữ liệu như ***h1***, ***div***, ***v.v*** và sau đó cùng các hàm tìm kiếm của BeautifulSoup để thu thập và lưu trữ vào database.

3.2 Tiền xử lí và làm sạch dữ liệu(NLP)

Bước tiền xử lí dữ liệu sẽ được mô tả tóm tắt như sau:

1. Làm sạch dữ liệu

Mục đích bước này là loại bỏ noise trong data. Đa phần noise là các thẻ HTML, JavaScript, và đương nhiên nếu cứ để noise để tiến hành xử lý sẽ dẫn đến kết quả xử lý không tốt.

Ví dụ đơn giản như sau:

Thông thường chúng ta hay loại bỏ noise là các thẻ HTML và JS như trên tuy nhiên thực tế noise có thể không chỉ là HTML, JS, cũng có thể là những cụm từ không cần thiết, hay ký tự không có ý nghĩa ("***\$%&##***").

Với các trường hợp thông thường, cách đơn giản và phổ biến nhất là sử dụng filter theo regex. Một đoạn code sử dụng regex trong python để xóa các kí tự đặc biệt như sau:

```
text = re.sub("&lt;/?.*?&gt;", " &lt;&gt; ", text)
```

2. Tách từ

Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất.

Ví dụ : từ “đất nước” được tạo ra từ 2 âm tiết “đất” và “nước”, cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành 1 bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động, ...

Tách từ chính xác hay không là công việc rất quan trọng, nếu không chính xác rất có thể dẫn đến việc ý nghĩa của câu sai, ảnh hưởng đến tính chính xác và truyền tải thông tin của dữ liệu.

3. Loại bỏ 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: đề, này, kia... Tiếng anh là những từ như: is, that, this...

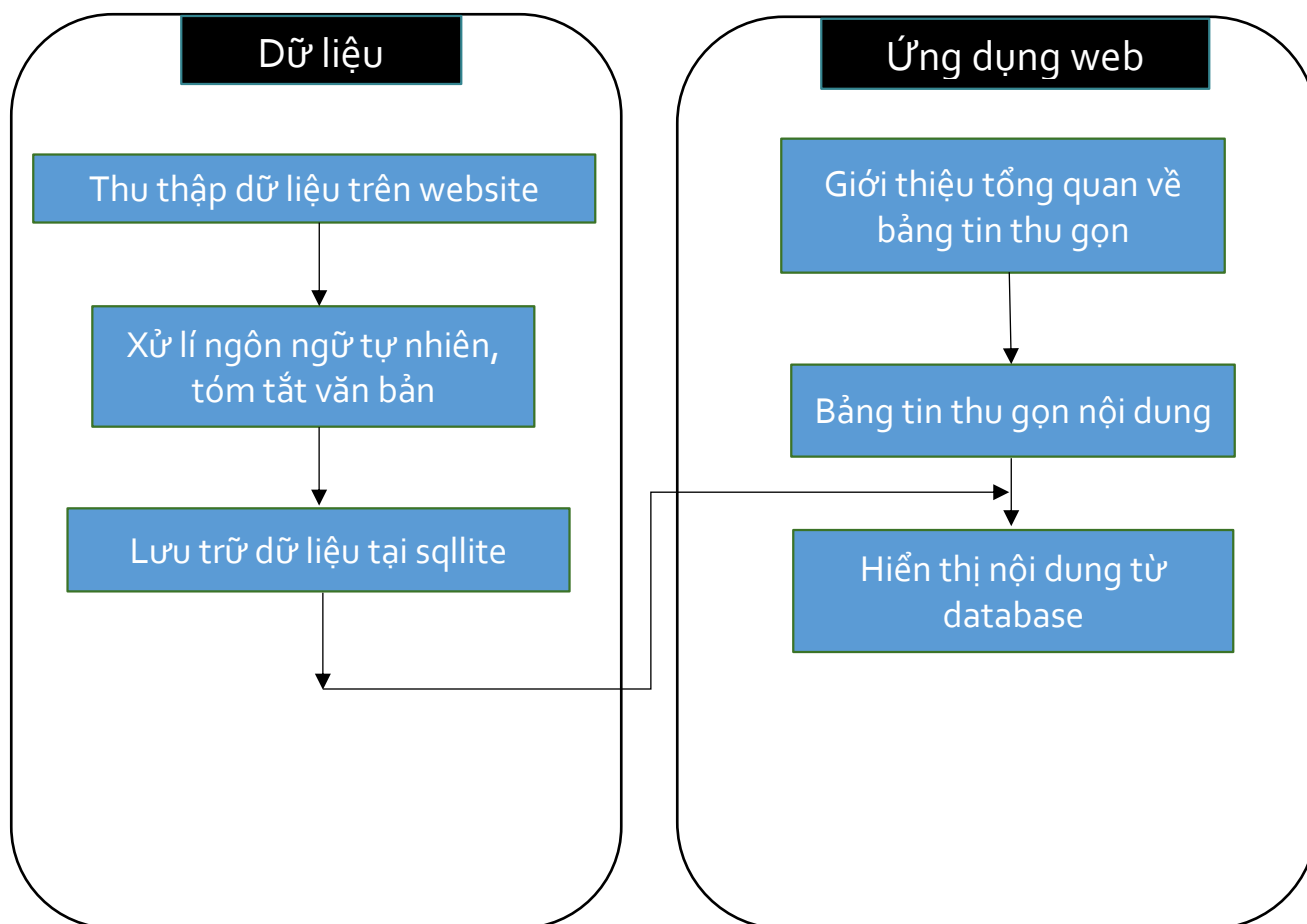
Một trong các phương pháp phổ biến để loại bỏ stopwords là dùng từ điển(chứa các stopwords) hoặc dựa vào tần suất xuất hiện của từ để xác định từ đó có phải stopwords hay không. Trong ứng dụng bản tin thu gọn này sử dụng phương pháp loại bỏ stopwords bằng từ điển, các stopwords được lưu trữ trong file *stopword_vietnam.txt* được cộng đồng Việt Nam đóng góp và sử dụng.

4. Rút gọn văn bản

Mục đích của việc trích xuất từ khóa tự động là tìm ra các cụm từ mô tả văn bản tốt nhất. Rút trích từ khóa là chuỗi của một hoặc nhiều từ vựng được rút ra từ văn bản. Văn bản được rút gọn vẫn đảm bảo nội dung, ý nghĩa so với nội dung của văn bản ban đầu. Sử dụng phương pháp Textrank và thư viện Gesim trong python để thực hiện công việc trích xuất từ khóa, rút gọn nội dung văn bản.

3.3 Mô hình tổng quát

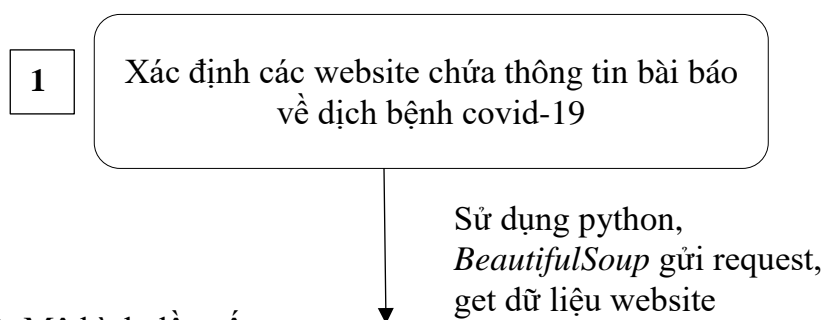
Sau đây là mô hình tổng quan của bảng tin thu gọn, bao gồm 2 phần chính là: phần dữ liệu và ứng dụng web.

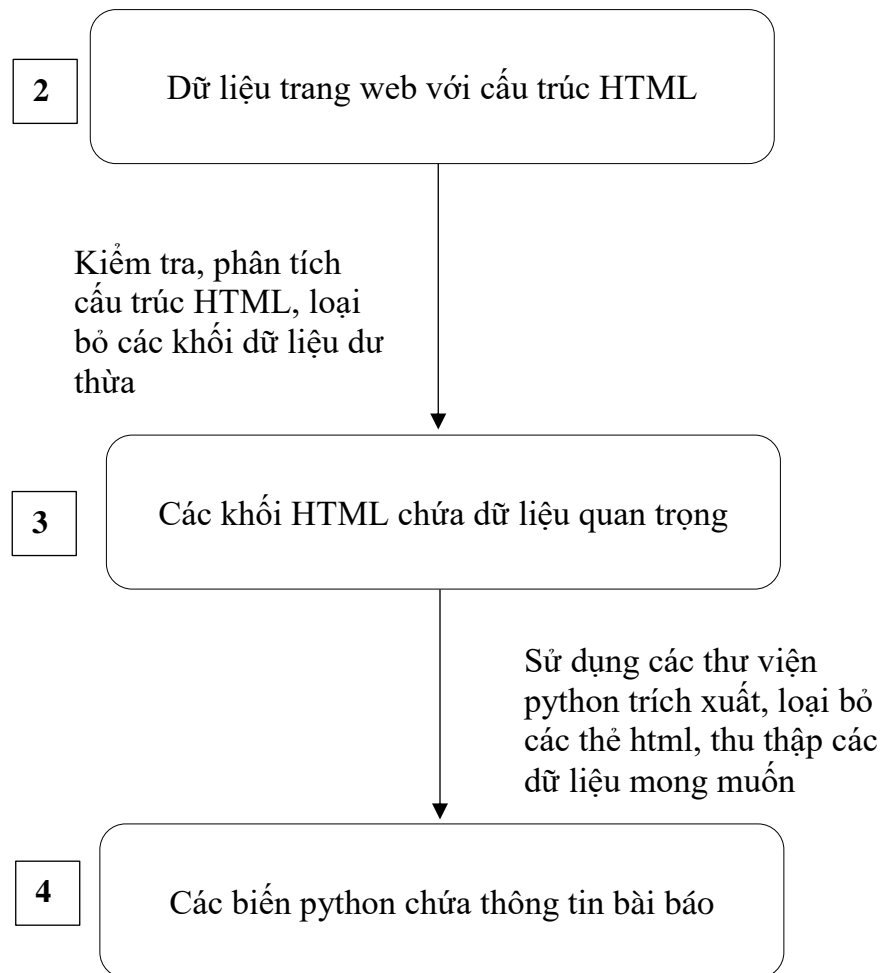


Hình 3.1 Mô tả tổng quát ứng dụng

3.3.1 Mô hình tổng quát thu thập dữ liệu

Sau đây là mô hình tổng quát quá trình thu thập dữ liệu:

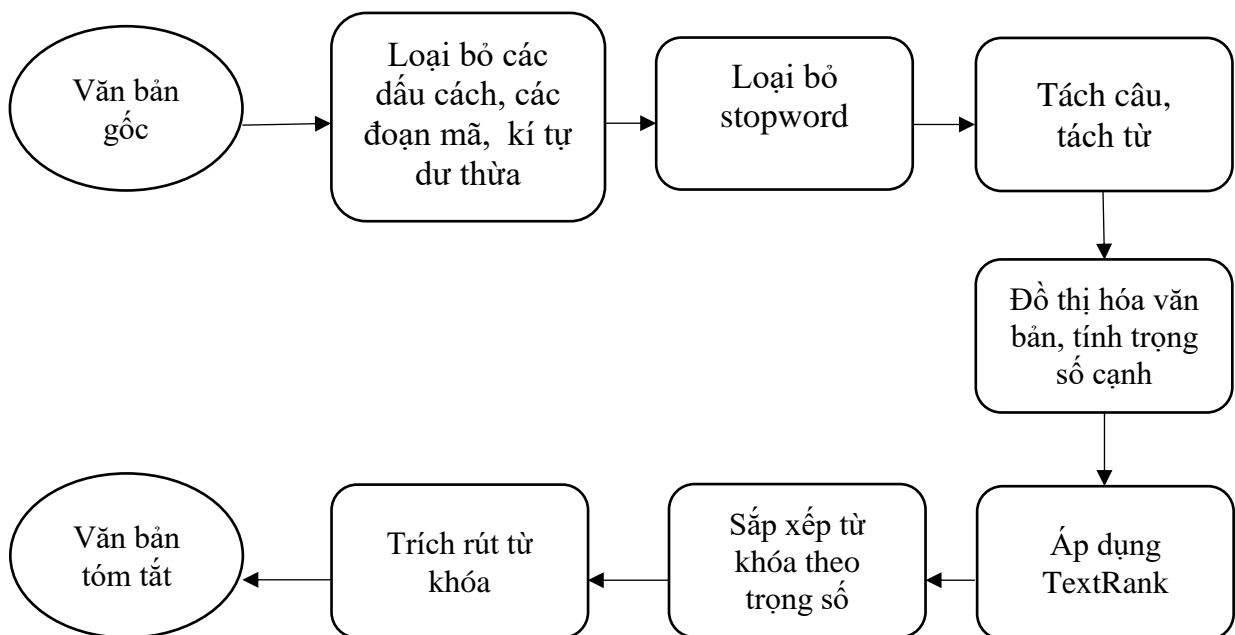




3.3.2 Tổng quát quá trình xử lý ngôn ngữ tự nhiên

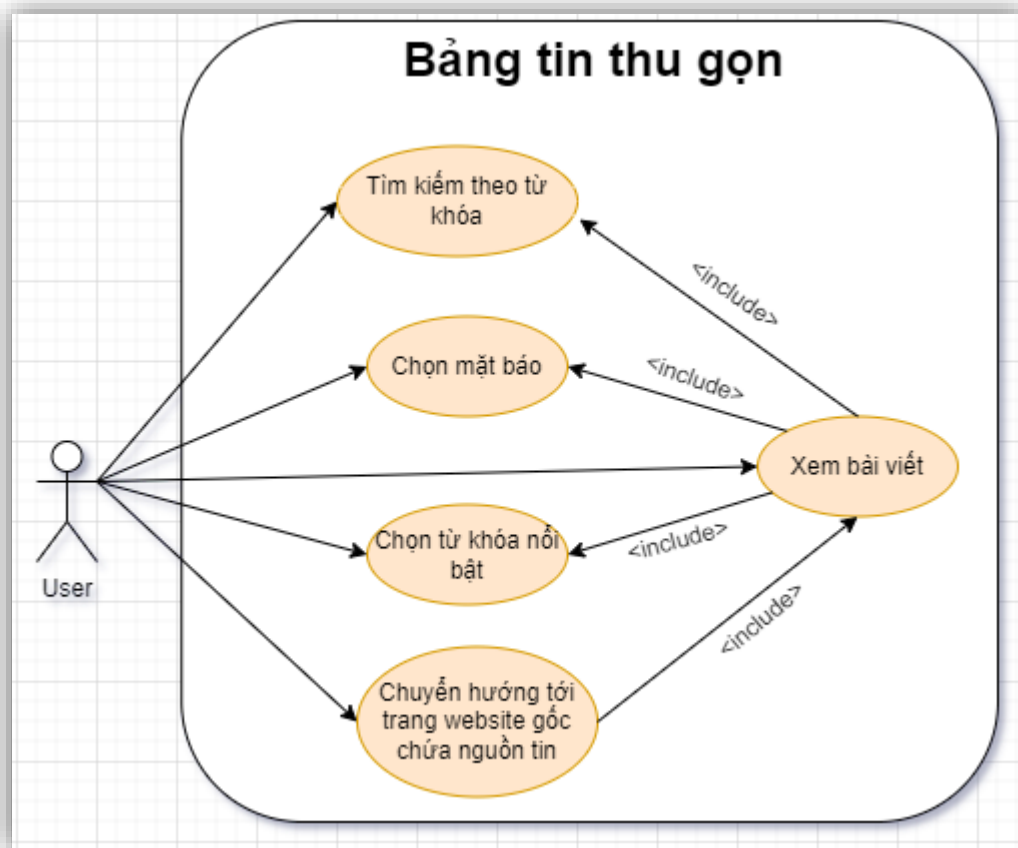
Sau khi có các dữ liệu thông tin bài báo ở dạng thô(chưa được tiền xử lí, làm sạch), cần phải xử lí và loại bỏ các loại bỏ noise là các thẻ HTML và JS, cũng có thể là những cụm từ không cần thiết, hay ký tự không có ý nghĩa (\$%&##"), hoặc cũng có thể là các stopwords. Tiếp theo là tách các từ Tiếng Việt, sau đó sử dụng thuật toán tóm tắt văn bản để trích xuất các nội dung trọng yếu, quan trọng của văn bản.

Sau đây là quá trình xử lí dữ liệu:



3.3.3 Use Case của bảng tin thu gọn

- Use case của người dùng:

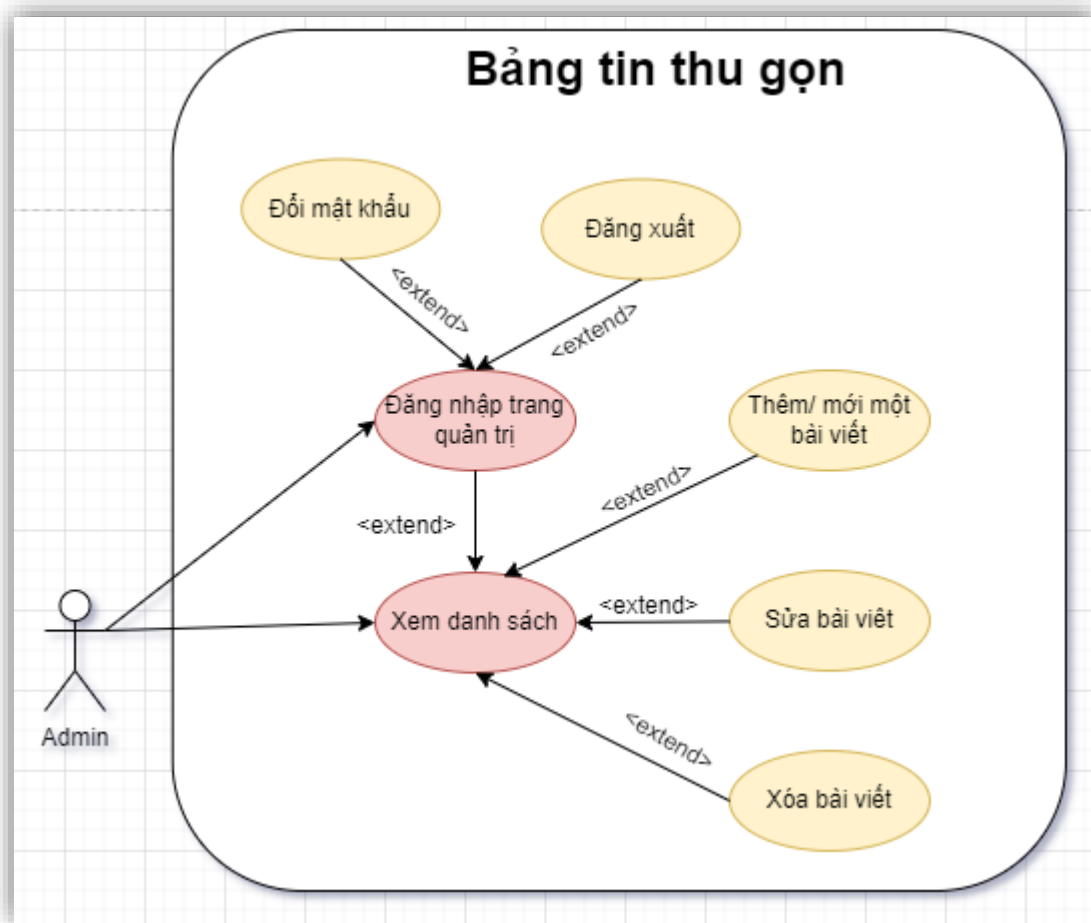


Hình 3.2 Use case của người dùng

Khi người dùng truy cập vào ứng dụng, User sẽ có các usecase cơ bản như sau:

- Click xem một bài viết bất kì
- Tìm kiếm bài viết theo từ khóa
- Lọc bài báo theo từ khóa nổi bật
- Lọc bài viết theo mặt báo

- Use case của Admin(quản trị viên, biên tập viên):



Hình 3.3 Use case của Admin

Admin muốn vào trang quản trị phải đăng nhập được vào hệ thống.

Các chức năng cơ bản của hệ thống đối với admin như sau:

- Xem danh sách các bài viết
- Thêm mới một bài viết
- Sửa một bài viết
- Xóa bài viết

CHƯƠNG 4: THỰC NGHIỆM

4.1 Các công nghệ và thư viện sử dụng

Để giải quyết bài toán thu thập, xử lý dữ liệu, tóm tắt văn bản, lưu trữ và tạo bảng tin thu gọn về dịch bệnh covid-19, đề tài sử dụng các công nghệ sau:

- Các nền tảng và thư viện:
 - +Ngôn ngữ lập trình: Python
 - + Các thư viện: Django Framework, Bootstrap, Gesim, Regex, . . .
- Công cụ sử dụng: Google Chrome, Pycharm
- Nền tảng lưu trữ dữ liệu: Sqlite3

4.2 Mô tả dữ liệu thu thập

Để tiến hành thực nghiệm, dữ liệu cần liên quan tới dịch bệnh covid-19, tiêm vắc xin covid-19, các hoạt động phòng, chống, hỗ trợ liên quan tới dịch bệnh covid-19. Dữ liệu sẽ được thu thập tại trang báo ***dantri.com.vn*** và ***baobinhduong.vn***

Các trường dữ liệu được lưu trữ thành 2 bảng tương ứng với 2 class trong Models Django như sau:

```
class covid_links_news(models.Model):
    link = models.CharField(max_length=500)
    tag_news = models.CharField(max_length=100, default="dantri.com.vn")
    def __str__(self):
        return self.link

class covid_news(models.Model):
    title = models.CharField(max_length=500)
    description = models.CharField(max_length=5000)
    content = models.CharField(max_length=5000)
    keywords = models.CharField(max_length=5000)
    url_news = models.CharField(max_length=5000)
    url_img = models.CharField(max_length=5000)
    publish_date = models.CharField(max_length=5000)
    summarize_content = models.CharField(max_length=5000, null=True)
    tag_news = models.CharField(max_length=100, default="dantri.com.vn")
    def __str__(self):
        return self.title
```

Hình 4.1 Các thuộc tính trong bảng dữ liệu

- Class *covid_links_news*:

Bảng 1: Mô tả bảng dữ liệu covid_links_news

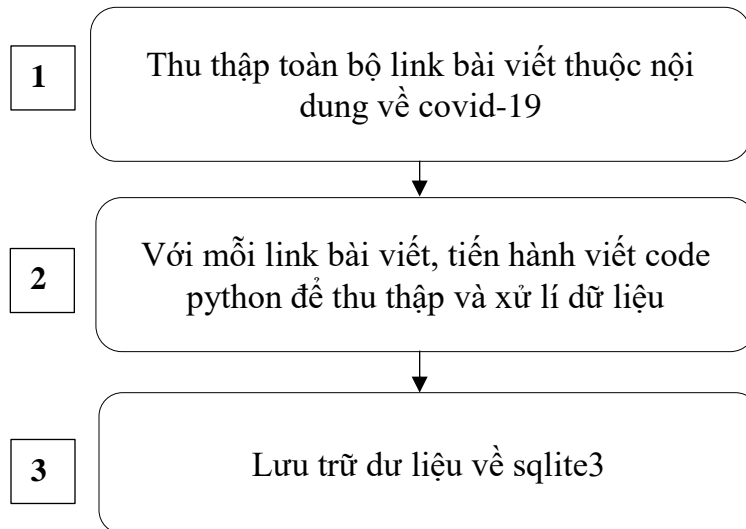
STT	Tên thuộc tính	Kiểu dữ liệu	Giải thích
1	link	Text	Link gốc bài báo covid-19
2	tag_news	Text	Chứa tên nguồn gốc bài báo. Giá trị mặc định là dantri.com.vn

- Class *covid_news*:

Bảng 2: Mô tả bảng dữ liệu covid_news

STT	Tên thuộc tính	Kiểu dữ liệu	Giải thích
1	title	Text	Tiêu đề tài báo
2	description	Text	Mô tả bài báo đã được xử lý và làm sạch
3	content	Text	Nội dung bài báo đã được xử lý và làm sạch
4	keywords	Text	Từ khóa của bài báo
5	url_news	Text	Link gốc bài báo
6	url_img	Text	Link ảnh bài báo
7	publish_date	Text	Ngày xuất bản bài báo
8	summarize_content	Text	Nội dung bài báo đã được rút gọn
9	tag_news	Text	Chứa tên nguồn gốc bài báo. Giá trị mặc định là dantri.com.vn

- Các bước thu thập và xử lý dữ liệu như sau:



4.3 Xây dựng bảng tin thu gọn về dịch bệnh Covid-19(sử dụng Django Framework)

4.3.1 Khởi tạo project Django

- Tạo project *hetuvan_vieclam* bằng Django trên Pycharm trên Terminal:

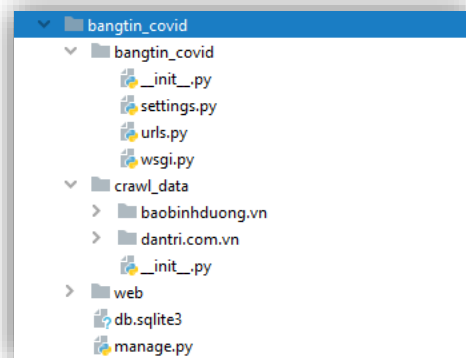
```
> django-admin startproject bangtin_covid
```

- Tạo thư mục *crawl_data* với 2 packpage là *dantri.com.vn* và *baobinhduong.vn*

- Tạo 1 app tên *web*

```
> python manage.py startapp Web
```

- Cấu trúc project được tạo như sau:



Hình 4.2 Cấu trúc project

Cấu trúc project gồm 3 phần chính:

+ Thư mục `bangtin_covid`: chứa các file cài đặt cũng như khai báo đường dẫn của project

+ `crawl_data`: chứa 2 packpage *dantri.com.vn* và *baobinhduong.vn*

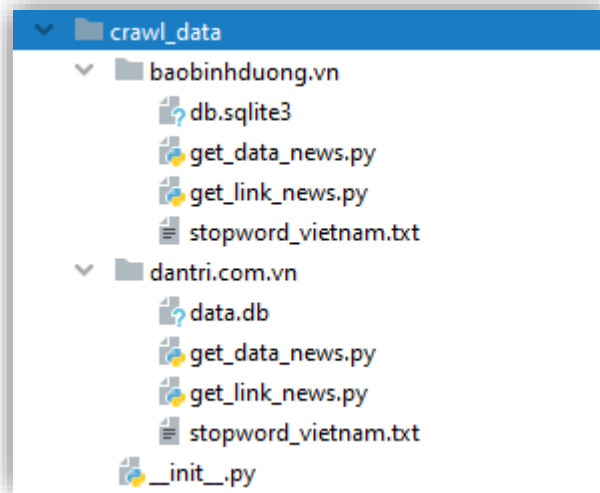
+ `web`: chứa các thư mục `static`, `templates`, `migrations`, và các file python:

- `models.py` dùng để khởi tạo cơ sở dữ liệu,
- `views.py` dùng để render file HTML, cũng như truyền, nhận dữ liệu từ front-end và xử lý dữ liệu thông qua `models.py`
- `apps.py`: khai báo tên apps ở bước khởi tạo ban đầu
- `admin.py`: khai báo các class có trong `models.py` để đăng kí chúng trên phần quản trị front-end
- `urls.py`: khai báo đường dẫn các hàm của file `views.py`, dùng để truyền dữ liệu từ front-end xuống `views.py` và `models.py`

+ `Filedb.sqlite3`: là database chứa dữ liệu của project

4.3.2 Thu thập dữ liệu covid-19

- Cấu trúc `crawl_data`:



Hình 4.3 Cấu trúc thư mục `crawl_data`

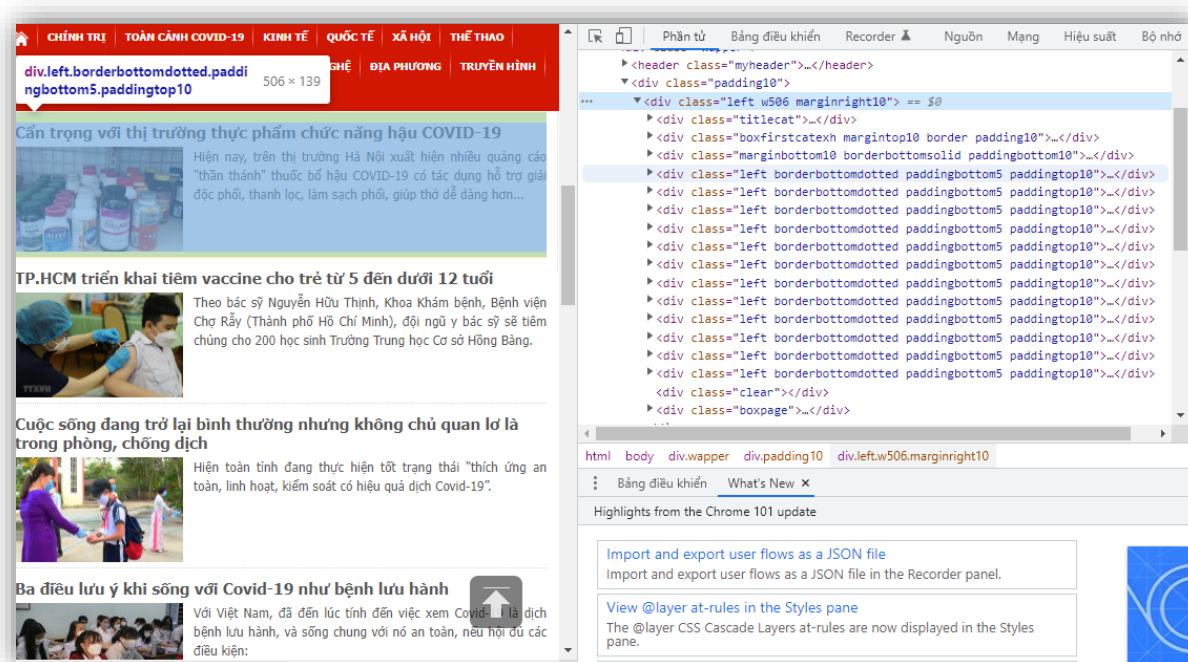
+ Mỗi packpage *dantri.com.vn* và *baobinhduong.vn* chứa 3 file:

- `get_link_news.py`: chứa code thu thập các link bài báo

Việc đầu tiên cần xác định các link chứa bài báo nằm tại thẻ nào của khối HTML, sau đó xác định class của chúng và tiến hành thu thập

Tiếp theo là sử dụng python và thư viện beautifulsoup để tách các khối HTML và dùng vòng lặp for để thu thập các link chứa bài báo

Cuối cùng là lưu link bài báo vào database



Hình 4.4 phân tích cấu trúc HTML trang chứa link bài báo

Tiến hành code python để get link bài báo về:

Hàm thu thập các link bài báo đối với **baobinhduong.vn**

#File `get_link_news.py`

```
def get_link_news_by_page(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.content, "html.parser")
    news_block = soup.findAll('div', class_="left
borderbottomdotted paddingbottom5 paddingtop10")
    link_news = [div.find('a').attrs['href']
for div in news_block]
    return link_news
```

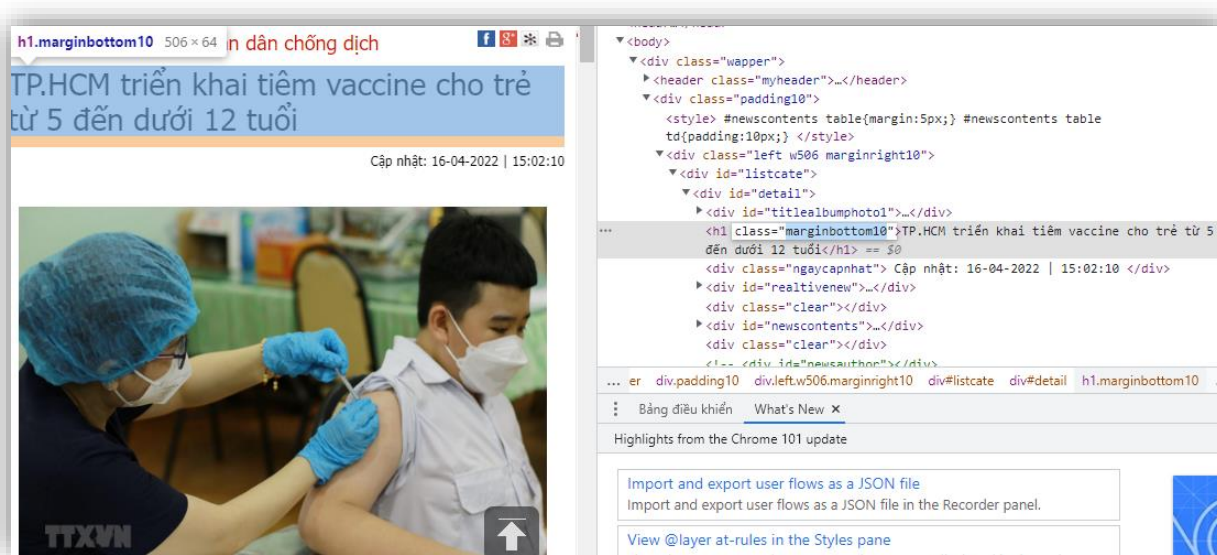
Hàm lưu trữ các link đã thu thập vào database:

#File *get_link_news.py*

```
def add_data(data):
    conn = sqlite3.connect(r"D:\Dai Hoc Thu Dau Mot\Nam 4\HK
    II\Doantotnghiep\bangtin_covid\db.sqlite3")
    query = "INSERT INTO web_covid_links_news(link,tag_news)
    VALUES (?,?)"
    tag_news = "baobinhduong.vn"
    for link in data:
        conn.execute(query, (link,tag_news))
        conn.commit()
        print("added to database ", link)
```

- *get_data_news.py*: chứa code thu thập dữ liệu đối với mỗi link bài báo thu thập được ở *get_link_news.py*

Tương tự các bước phân tích các khối HTML, với mỗi link bài báo tiến hành tìm ra các khối, class chứa thông tin bài báo, sử dụng python và các thư viện cần thiết để tiến hành thu thập và lưu trữ vào database.



Hình 4.5 Phân tích cấu trúc HTML của bài báo

Việc thu thập dữ liệu cần gắn liền với việc xử lý ngôn ngữ tự nhiên, sau khi thu thập được dữ liệu, tiến hành xử lý ngôn ngữ cho dữ liệu như sau:

#File *get_link_news.py*

```
def text_prossesing(text):
    text = re.sub(r"\s|s+", ' ', text.strip())
    # load stopword
    stopwords_vietnam = []
    with open('stopword_vietnam.txt', 'r', encoding="utf8")
        as f:
            for line in f:
                stopwords_vietnam.append(line.strip())
    text = text.lower()
    # remove tags
    text = re.sub("</?.*?>", " &lt;&gt; ", text)
    # remove stopwords
    text = ' '.join([word for word in text.split() if word not in
        stopwords_vietnam])
    text = tokenize(text)
    return text
```

- *stopword_vietnam.txt*: tệp chứa các stopwords Tiếng Việt
- Tương tự với mặt báo ***dantri.com.vn***, tiến hành thu thập các link và dữ liệu covid-19 bằng python và các thư viện cần thiết.

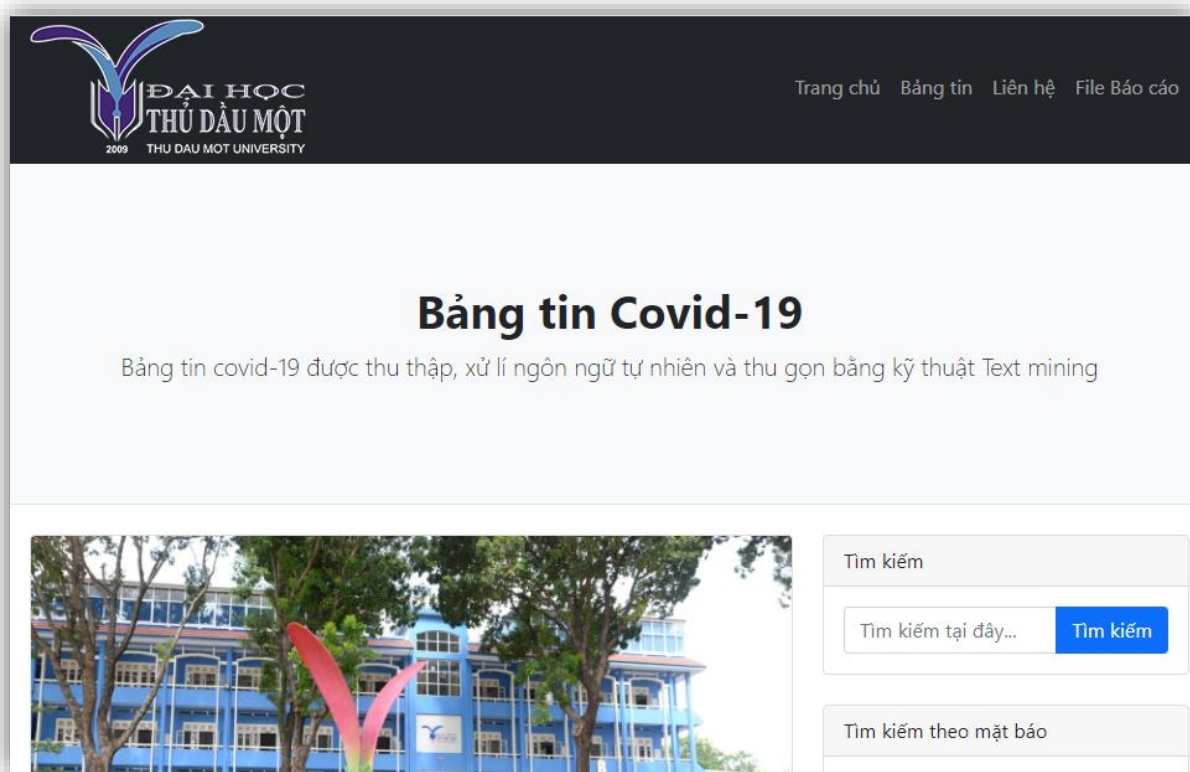
id	link	tag_news
1	https://baobinhduong.vn/can-trong-voi-thi-truong-thuc-pham-chuc-nang-hau-covid-19-a270078.html	baobinhduong.vn
2	https://baobinhduong.vn/tp-hcm-trien-khai-tiem-vaccine-cho-tre-tu-5-den-duoi-12-tuoi-a269391.html	baobinhduong.vn
3	https://baobinhduong.vn/cuoc-song-dang-tro-lai-binh-thuong-nhung-khong-chu-quan-lo-la-trong-phong-...	baobinhduong.vn
4	https://baobinhduong.vn/ba-dieu-luu-y-khi-song-voi-covid-19-nhu-benh-luu-hanh-a266716.html	baobinhduong.vn
5	https://baobinhduong.vn/nhung-sai-lam-khi-su-dung-test-nhanh-covid-19-a266598.html	baobinhduong.vn
6	https://baobinhduong.vn/binh-duong-ghi-nhan-them-truong-hop-nhap-can-hiem-omicron-a265358.html	baobinhduong.vn
7	https://baobinhduong.vn/chu-dong-phong-chong-dich-covid-19-dac-biet-la-bien-the-moi-cua-vi-rut-sars-	baobinhduong.vn
8	https://baobinhduong.vn/xuat-hien-cac-o-dich-phuc-tap-nam-dinh-tap-trung-chong-dich-dip-tet-a264664	baobinhduong.vn
9	https://baobinhduong.vn/thanh-hoa-ghi-nhan-so-ca-mac-covid-19-cao-nhat-tu-truoc-toi-nay-a264531.html	baobinhduong.vn
10	https://baobinhduong.vn/covid-19-thanh-pho-can-tho-tro-thanh-vung-xanh-tu-ngay-24-1-a264300.html	baobinhduong.vn

Hình 4.6 Dữ liệu sau khi thu thập

id	title	description	content	keywords	url_news	url_img	publish_date	summarize_content	tag_news
35	Hơn 20.000 ca C...	Ngày 15/4, Bộ Y ...	tính 16h ngày 14 ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ sáu, 15/0...	tính 16h ngày 14 / 4 ...	dantri.com.vn
36	Mỹ cấp phép bộ ...	Cục Quản lý Thự...	một điểm xét ng...	xét nghiệm Cov...	https://dantri.com.vn/t...	http://icdn.d...	Thứ sáu, 15/0...	hãng inspectr được ...	dantri.com.vn
37	Hà Nội dự kiến t...	Mục tiêu thành ...	hà_nội dự_kiến t...	vaccine hà nội	https://dantri.com.vn/s...	http://icdn.d...	Thứ sáu, 15/0...	hà_nội dự_kiến tiêm ...	dantri.com.vn
38	Hà Nội thêm 1.6...	Tối 14/4, Sở Y tế ...	trong 24 giờ qua , ...	Covid -19 hà nội	https://dantri.com.vn/s...	http://icdn.d...	Thứ năm, 14/...	một_số quận , huyện...	dantri.com.vn
39	Hơn 23.000 ca C...	Ngày 14/4, Bộ Y ...	tính 16h ngày 13 ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ năm, 14/...	gần 200 trẻ dưới 12 t...	dantri.com.vn
40	WHO cảnh báo ...	Ủy ban khẩn cấp...	một phụ_nữ xét...	Covid -19 đại di...	https://dantri.com.vn/t...	http://icdn.d...	Thứ năm, 14/...	đại_dịch covid - 19 c...	dantri.com.vn
41	Biến thể phụ ch...	Hai biến thể phụ...	người xét_nghiệ...	dịch Covid -19 ...	https://dantri.com.vn/t...	http://icdn.d...	Thứ năm, 14/...	theo cơ_quan y_tế ba...	dantri.com.vn
42	Cả nước có 24.6...	Ngày 13/4, Bộ Y ...	tính 16h ngày 12 ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ tư, 13/04/...	tính 16h ngày 12 / 4 ...	dantri.com.vn
44	Việt Nam xây dự...	Việt Nam sẽ xây ...	gs . ts phan trọng...	dịch Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ tư, 13/04/...	ts phan trọng lần , cụ...	dantri.com.vn
45	Việt Nam chưa g...	Theo Cục Y tế Đ...	đại diện cục y_tế ...	biến thể phụ bi...	https://dantri.com.vn/s...	http://icdn.d...	Thứ tư, 13/04/...	đại diện cục y_tế dự...	dantri.com.vn
46	22.804 F0 mới tr...	Ngày 12/4, Bộ Y ...	tính 16h ngày 11 ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ ba, 12/04/...	tính 16h ngày 11 / 4 ...	dantri.com.vn
47	Hà Nội ghi nhận...	Tối 12/4, Sở Y tế ...	trong 24 giờ qua , ...	hà nội Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ ba, 12/04/...	một_số quận , huyện...	dantri.com.vn
48	Ca Covid-19 tăn...	Gửi chức thành ...	trẻ_em tiêm vacci...	vaccine Covid-...	https://dantri.com.vn/t...	http://icdn.d...	Thứ ba, 12/04/...	quyết_định trên hiệu...	dantri.com.vn
49	Triệu chứng càn...	Tâm lý căng thả...	6 triệu_chúng cá...	Covid -19 trăm ...	https://dantri.com.vn/s...	http://icdn.d...	Thứ ba, 12/04/...	6 triệu_chúng cảnh_...	dantri.com.vn
50	Cả nước có hơn ...	Ngày 11/4, Bộ Y ...	tính 16h ngày 10 ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ hai, 11/04/...	tính 16h ngày 10 / 4 ...	dantri.com.vn
51	Hà Nội thêm 2.0...	Tối 11/4, Sở Y tế ...	24 giờ qua , hà_n...	hà nội Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Thứ hai, 11/04/...	một_số quận , huyện...	dantri.com.vn
52	Hà Nội phát hiệ...	Tối 10/4, Sở Y tế ...	trong 24 giờ qua , ...	hà nội F0 Covid...	https://dantri.com.vn/s...	http://icdn.d...	Chủ nhật, 10/...	hà_nội dự_kiến quý ì...	dantri.com.vn
53	Ca mắc mới Cov...	Ngày 10/4, Bộ Y ...	tính 16h ngày 9 / ...	Covid -19	https://dantri.com.vn/s...	http://icdn.d...	Chủ nhật, 10/...	tính 16h ngày 9 / 4 đ...	dantri.com.vn

Hình 4.7 Dữ liệu bài báo sau khi thu thập và xử lý

4.3.3 Xây dựng giao diện ứng dụng

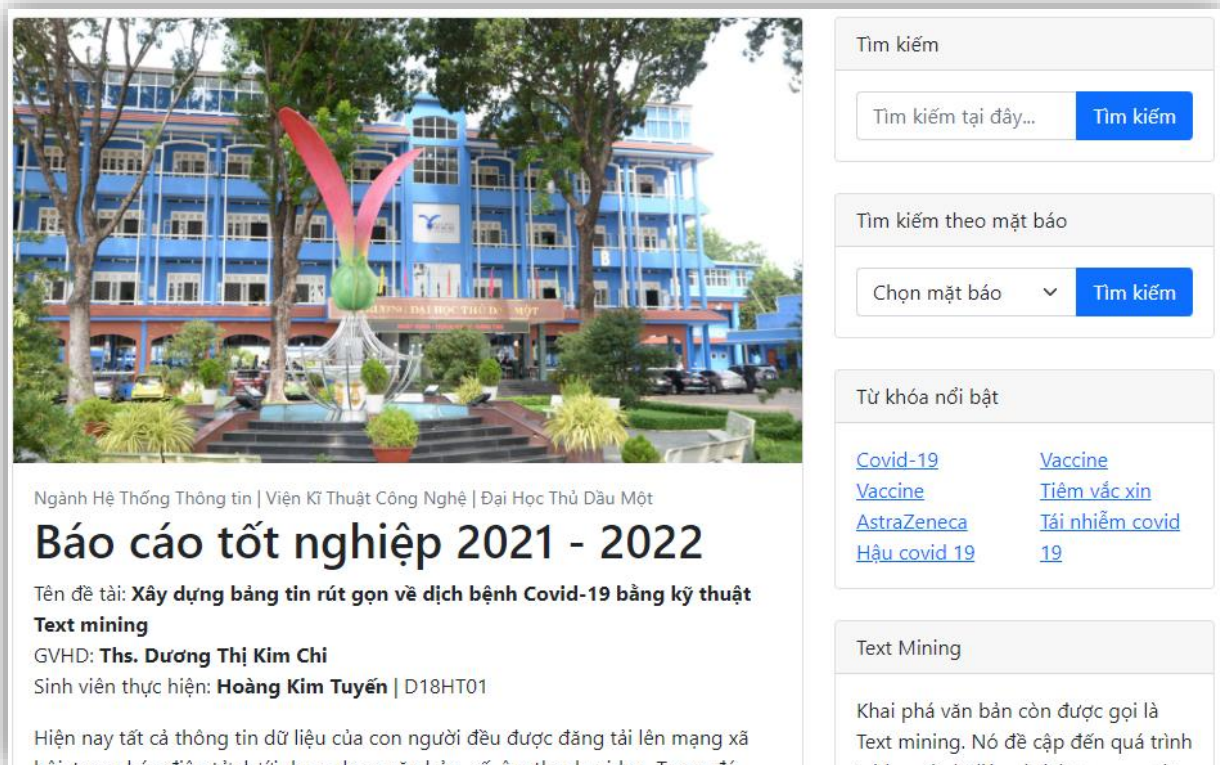


Hình 4.8 Trang chủ của ứng dụng

Bố cục chính gồm có:

- + Header: chứa logo, thanh menu gồm các page: trang chủ, bảng tin, liên hệ, file báo cáo
- + Body: chứa nội dung cụ thể của từng bài viết
- + Sidebar bên phải: chứa ô tìm kiếm, dropdown lựa chọn mặt báo, danh mục các từ khóa nổi bật
- + Footer: Copyright © Thu Dau Mot University 2022

- Trang chủ tìm kiếm giới thiệu thông tin về ứng dụng



Hình 4.9 Trang chủ tìm kiếm của ứng dụng

- Trang bảng tin chứa các mục bài báo như sau: ảnh thu nhỏ của bài báo, ngày xuất bản, tiêu đề, mô tả ngắn và nút xem chi tiết.



Hình 4.10 Bảng tin Covid-19

- Với mỗi trang bài viết chứa các nội dung sau:
 - + ảnh bài viết, ngày xuất bản, tiêu đề bài viết, từ khóa bài viết, link nguồn báo, mô tả bài viết, nội dung tóm tắt, button chuyển tới nguồn báo và các bài viết được gợi ý ngẫu nhiên của ứng dụng.
 - + Đối với mỗi bài viết sẽ có thống kê số liệu Covid được nhúng thông qua API của bộ y tế.

Đà Nẵng: Những người già yếu được tiêm vaccine COVID-19 tại nhà

Từ khóa: COVID-19

Nguồn báo: <https://baobinhduong.vn/da-n-ng-nhung-nguoi-gia-yeu-duoc-tiem-vaccine-covid-19-tai-nha-a261705.html>

Mô tả: Lãnh đạo thành phố Đà Nẵng chỉ đạo ngành y tế tổ chức điểm tiêm vaccine lưu động và đến tận nhà để tiêm cho những trường hợp là người lớn tuổi không thể di chuyển đến các điểm tiêm.

Nội dung tóm tắt: phát biểu chỉ đạo , phó chủ tịch ủy ban nhân thành phố thành phố đà nẵng ngỏ thị kim yển yêu cầu các đơn vị , địa phương , khi phát hiện ca mắc covid - 19 mới tại khu công nghiệp phải có cơ chế phối hợp chặt chẽ , hạn chế không để dịch bệnh lây lan đến các địa phương khác . qua các công nhân khu công nghiệp trên địa bàn quận liên chiều trở về nhà ghi nhận mắc covid - 19 , làm lây lan dịch bệnh trên địa bàn huyện hòa vang , phó chủ tịch ủy ban nhân thành phố đà nẵng cho rằng , các đơn vị , địa phương chưa đánh giá đúng mức độ nguy cơ của dịch bệnh . liên quan đến công tác tiêm vaccine , phó chủ tịch ủy ban nhân thành phố đà nẵng cho hay , theo số liệu thống kê của ngành y tế , còn một số trường hợp chưa được tiêm mũi 1 vaccine , dù những người này mong muốn được tiêm , đa số là những trường hợp lớn tuổi , khó khăn trong di chuyển .

[Xem chi tiết tại nguồn báo→](#)

Thống kê

Thống kê dịch Covid-19

Thế giới | **Việt Nam**

Ca nhiễm	Hôm nay	Tử vong
10.666.751	+ 0	43.049


Tỉnh / Thành Phố	Tổng số ca	Hôm nay	Tử vong
Bình Dương	320.589	+0	3.514
Phú Thọ	317.992	+0	97
Nam Định	294.706	+0	149

Hình 4.11 Giao diện bài viết thu gọn nội dung về Covid-19

Nội dung tóm tắt: tính 16h ngày 21 / 4 đến 16h ngày 22 / 4 , trên hệ thống quốc gia quản lý ca bệnh covid - 19 ghi nhận 11.160 ca nhiễm mới (giảm 869 ca ngày đó) tại 59 tỉnh , thành phố (có 8.015 ca trong cộng đồng) . trong đợt dịch thứ 4 (từ ngày 27 / 4 / 2021 đến nay) , số ca nhiễm ghi nhận trong nước là 10.536.576 ca , trong có 9.076.448 bệnh nhân đã được công bố khỏi bệnh . tổng số ca tử vong do covid - 19 tại việt nam tính đến nay là 42.998 ca , chiếm tỷ lệ 0,4 % tổng số ca nhiễm ; xếp thứ 24 / 227 vùng lãnh thổ .


[Xem chi tiết tại nguồn báo→](#)

Có thể bạn sẽ thích



Thứ ba, 19/04/2022 - 10:30

Cảnh báo: Covid-19 tấn công đường tiêu hóa khiến bé gái loét dạ dày



Chủ nhật, 17/04/2022 - 17:57

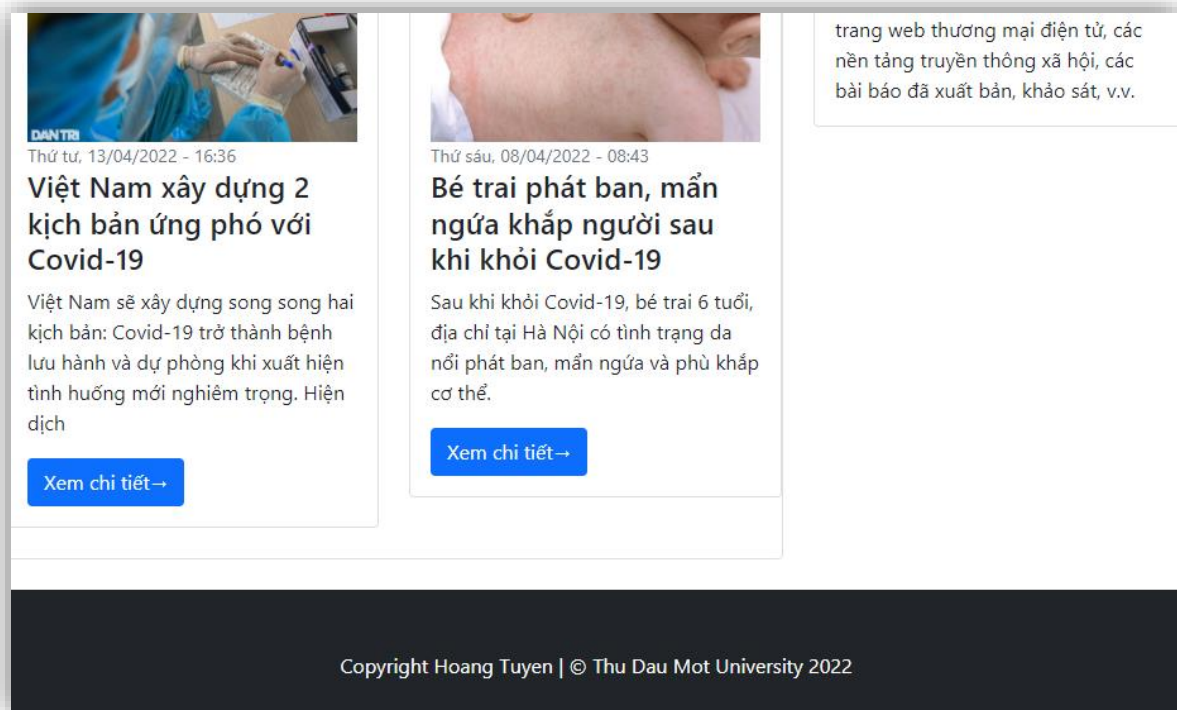
14.739 ca mắc mới, đã có 12.414 trẻ từ 5-11 tuổi được tiêm vaccine

Tỉnh / Thành Phố	Tổng số ca	Hôm nay	Tử vong
Hà Nội	1.590.311	+0	1.220
Hồ Chí Minh	610.296	+0	19.984
Nghệ An	482.297	+0	145
Bắc Giang	385.488	+0	91
Vĩnh Phúc	365.672	+0	19
	362.17		

Cập nhật lần cuối: 14:55 06/05/2022 Nguồn: Bộ Y Tế

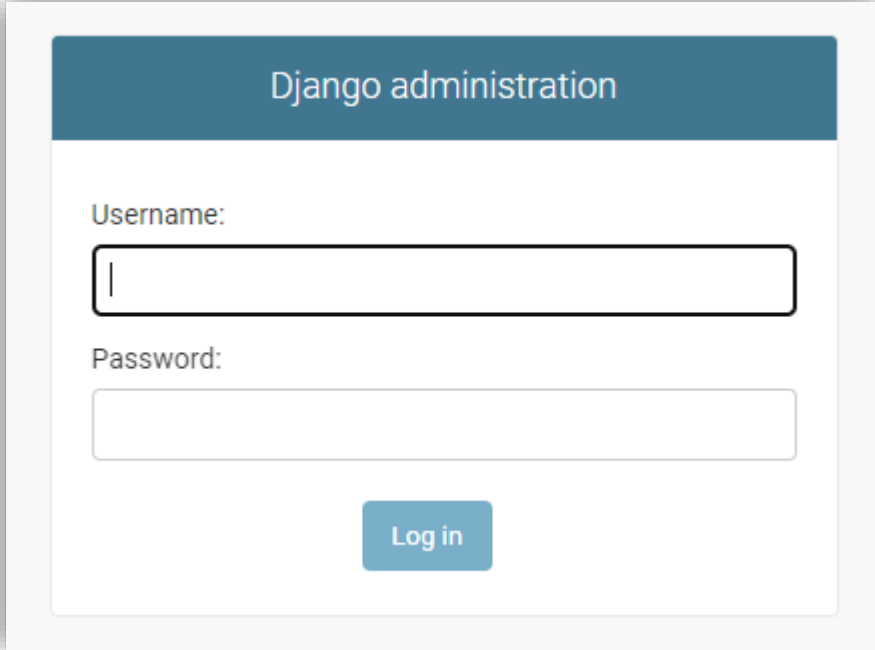
Hình 4.12 Giao diện phần dưới bài viết

- Phần footer chứa Copyright Đại Học Thủ Dầu Một:



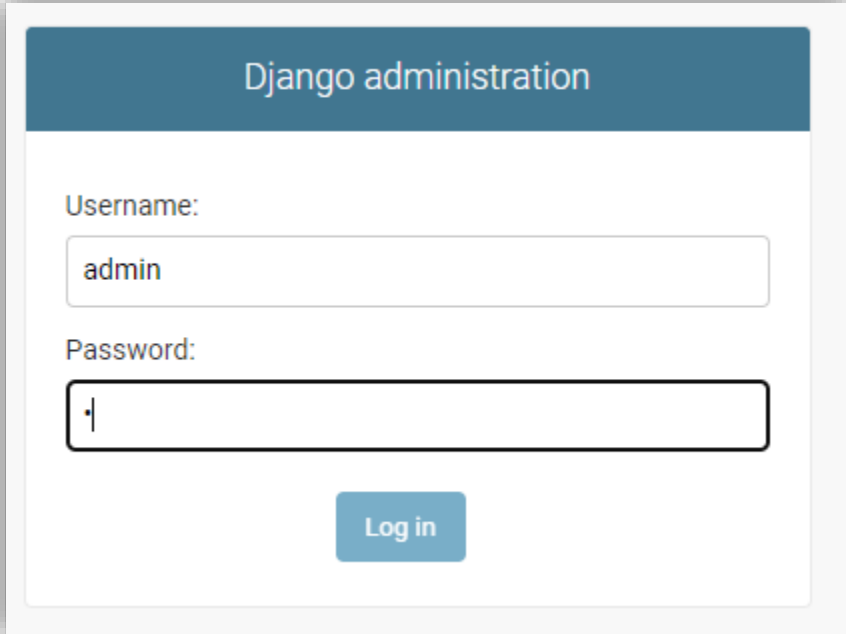
Hình 4.13 Giao diện footer của bảng tin

4.3.4 Giao diện và chức năng trang quản trị



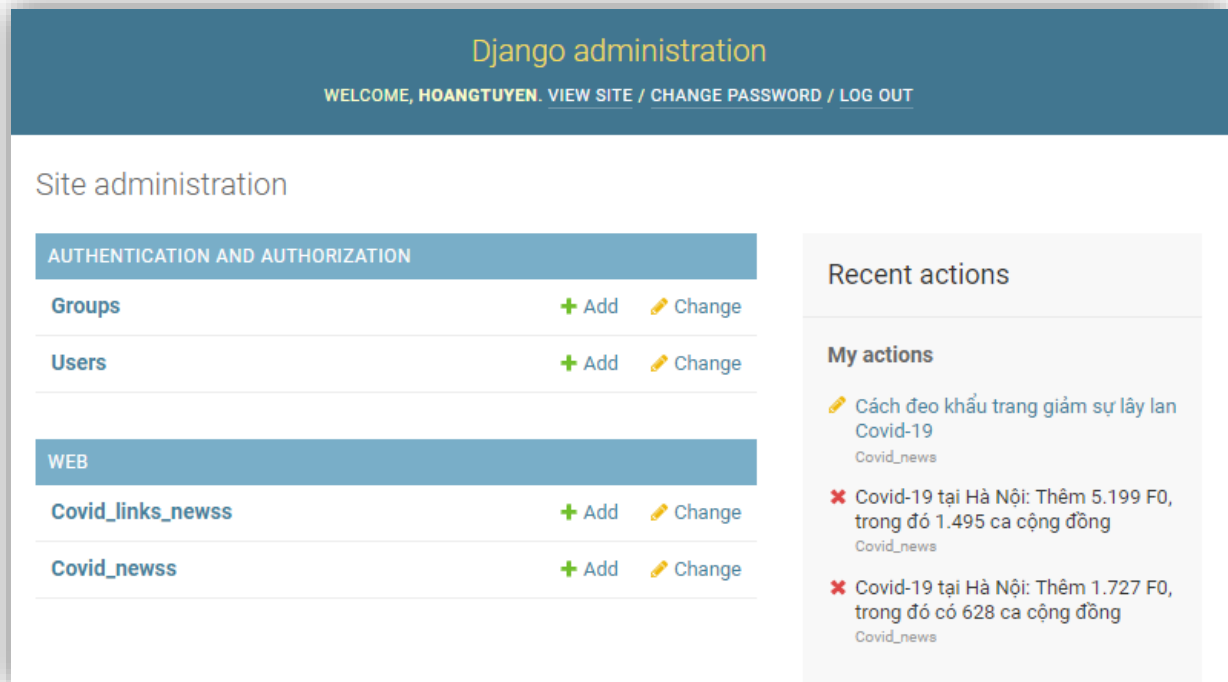
The screenshot shows the Django administration login interface. At the top, there is a dark blue header with the text "Django administration" in white. Below the header, the form has two input fields: "Username:" and "Password:". The "Username:" field is currently empty, and the "Password:" field is also empty. Below these fields is a blue button labeled "Log in".

Hình 4.14 Giao diện đăng nhập trang admin



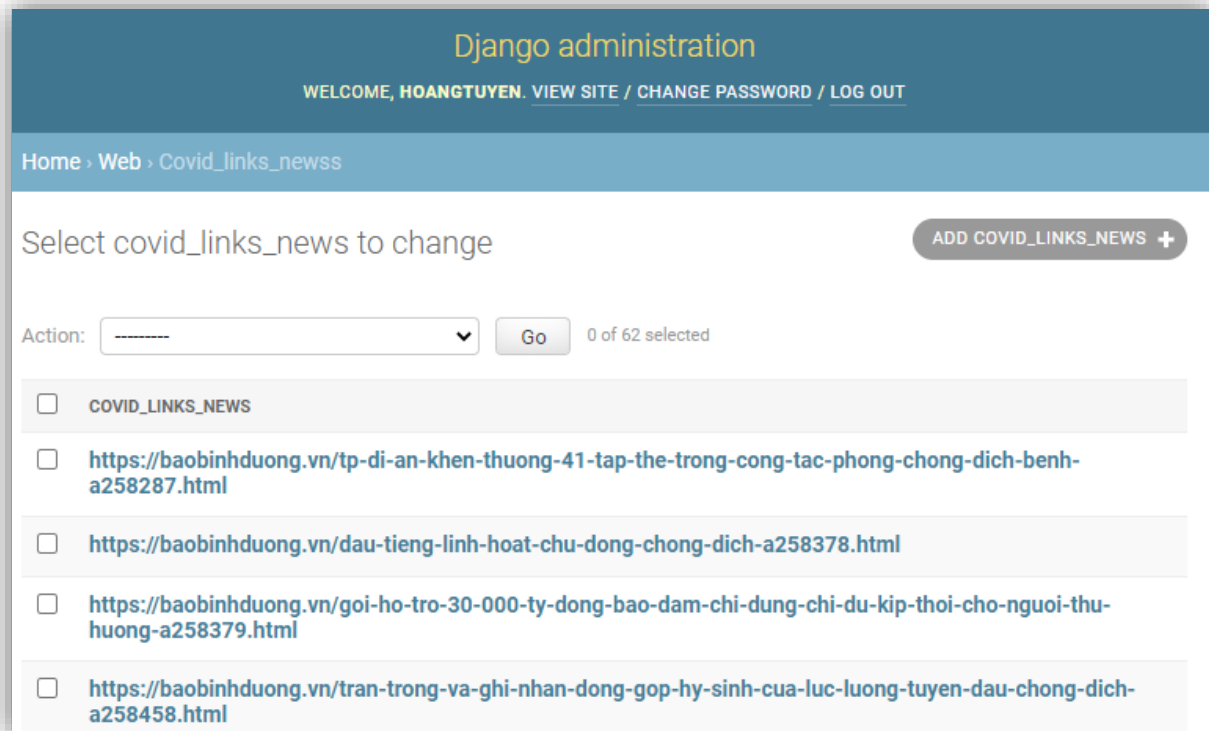
This screenshot shows the same Django administration login interface as the previous one, but with the username field filled with the text "admin". The password field remains empty. The "Log in" button is still present at the bottom.

Hình 4.15 Giao diện đăng nhập trang admin

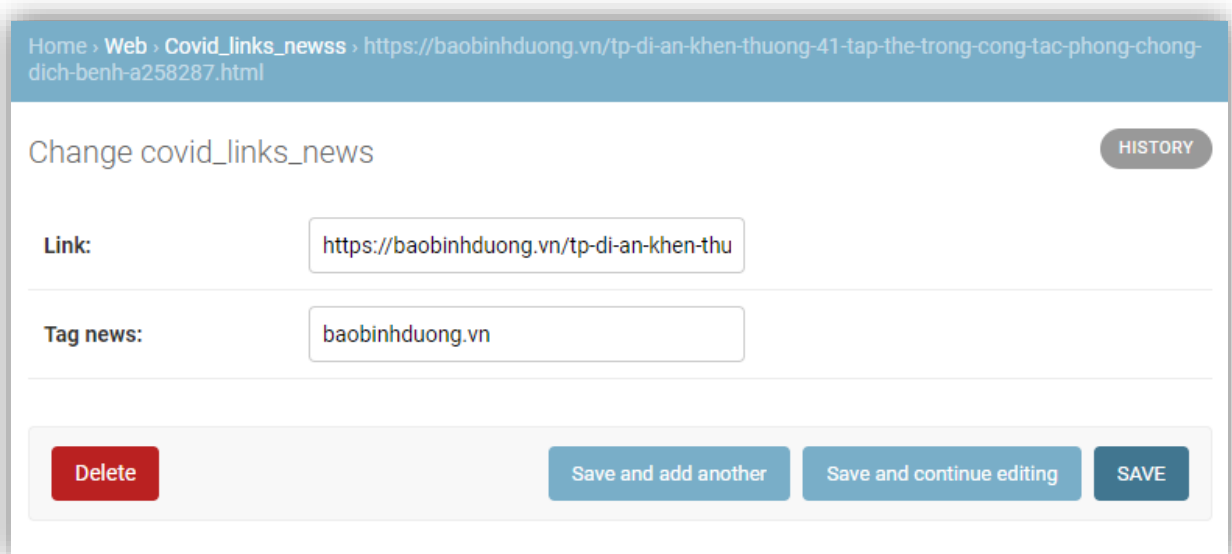


Hình 4.16 Giao diện trang quản trị

- Phần quản trị gồm 2 mục chính:
 - + Quản trị các user đăng nhập, phân quyền cho user có thể đăng nhập vào hệ thống
 - + Quản trị dữ liệu của ứng dụng: gồm 2 class được khai báo trong file *models.py* gồm *covid_links_news* và *covid_news*
- Quản trị *covid_links_news*:
 - + Chứa danh sách các link bài báo
 - + Chức năng thêm, xóa, sửa

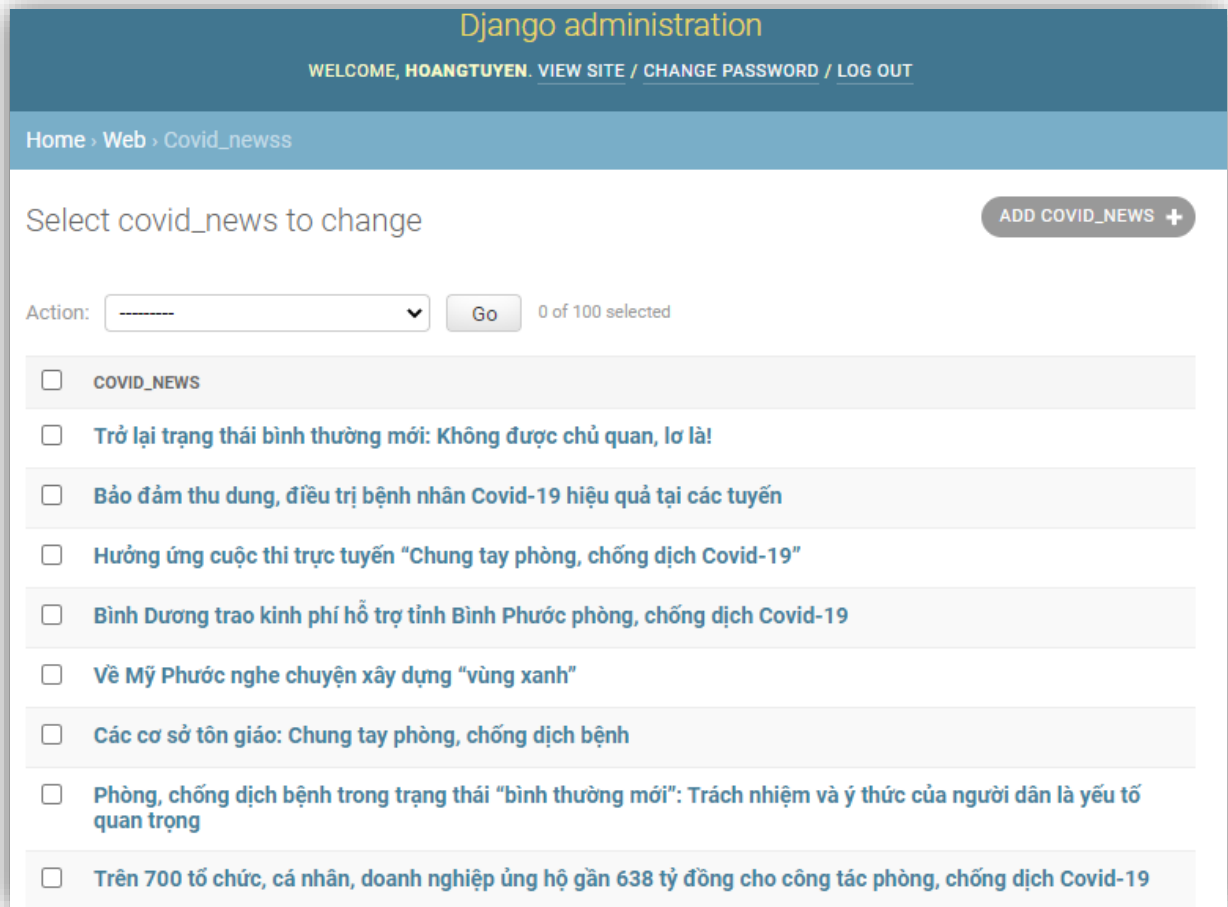


Hình 4.17 Giao diện quản trị link bài viết

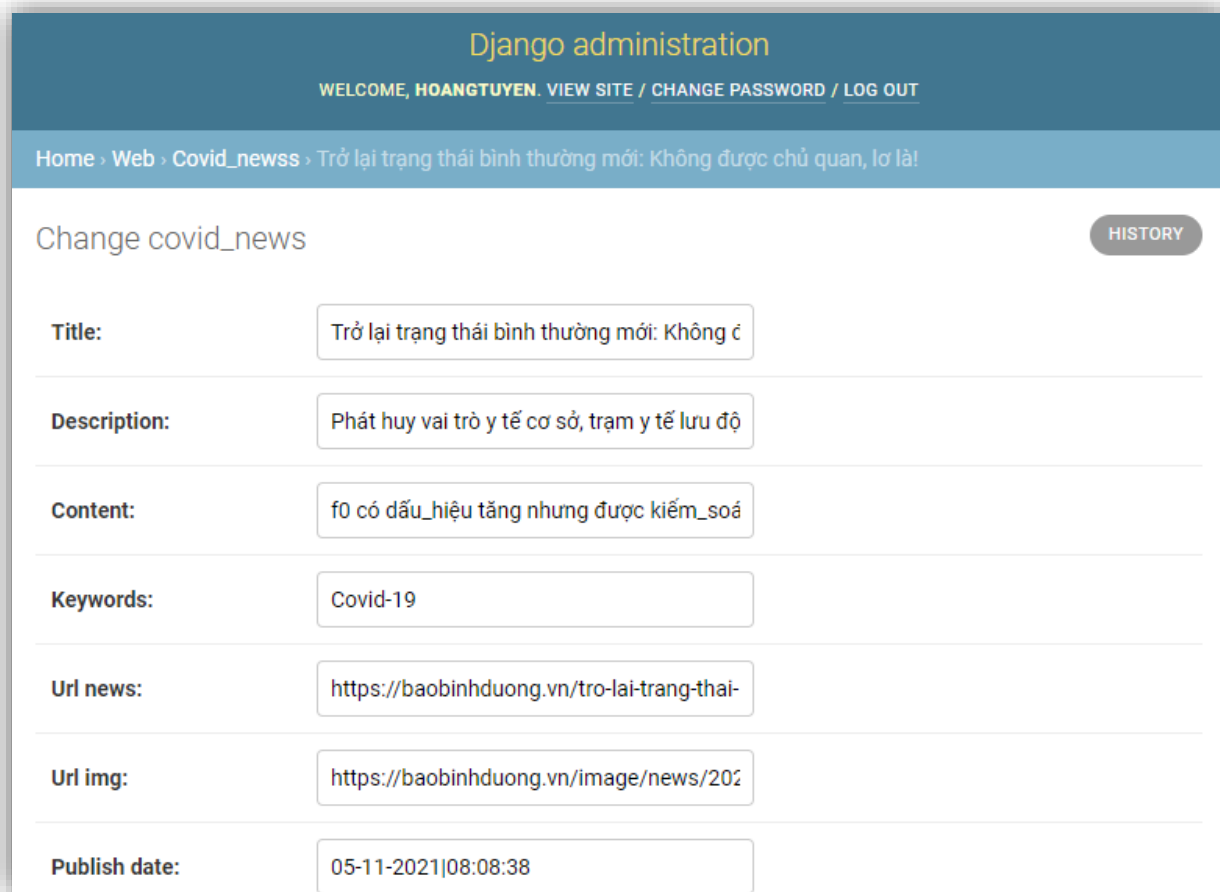


Hình 4.18 Giao diện thêm/xóa/sửa link bài viết

- Quản trị *covid_links_news*:
- + Chứa danh sách các bài báo
- + Chức năng thêm, xóa, sửa



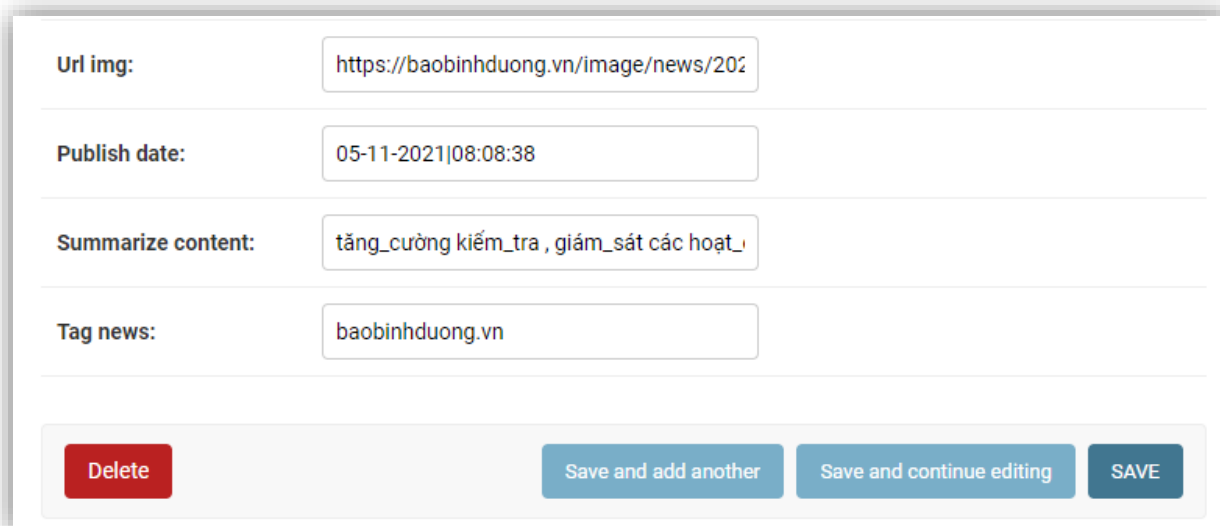
Hình 4.19 Giao diện quản trị các bài viết



The screenshot shows the Django administration interface for editing a news item. The header is dark blue with the text "Django administration" in yellow. Below the header, there is a navigation bar with links: "WELCOME, HOANGTUYEN. VIEW SITE / CHANGE PASSWORD / LOG OUT". The main content area has a breadcrumb trail: "Home > Web > Covid_newss > Trở lại trạng thái bình thường mới: Không được chủ quan, lơ là!". The title of the form is "Change covid_news" with a "HISTORY" button. The form fields are:

Title:	Trở lại trạng thái bình thường mới: Không c
Description:	Phát huy vai trò y tế cơ sở, trạm y tế lưu độ
Content:	f0 có dấu_hiệu tăng nhưng được kiểm_soá
Keywords:	Covid-19
Url news:	https://baobinhduong.vn/tro-lai-trang-thai-
Url img:	https://baobinhduong.vn/image/news/202
Publish date:	05-11-2021 08:08:38

Hình 4.20 Giao diện thêm/xóa/sửa một bài viết



The screenshot shows the bottom section of the Django administration interface for editing a news item. The form fields are:

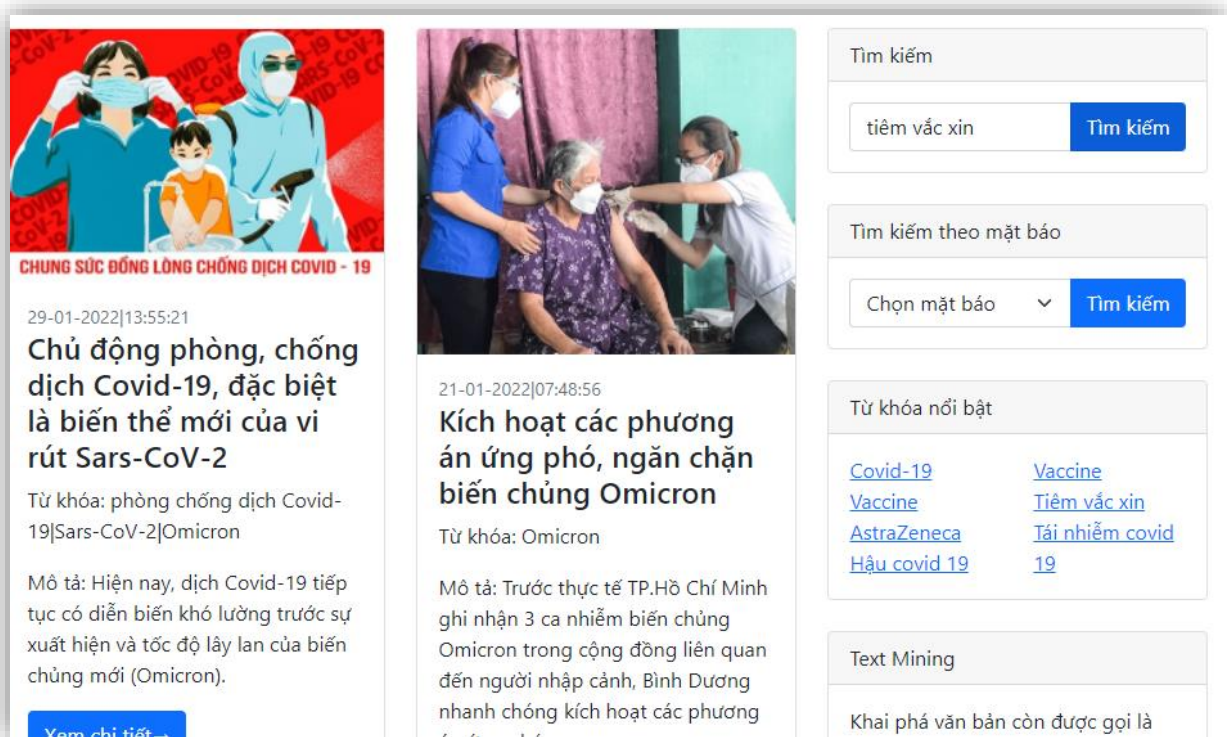
Url img:	https://baobinhduong.vn/image/news/202
Publish date:	05-11-2021 08:08:38
Summarize content:	tăng_cường kiểm_tra , giám_sát các hoạt_
Tag news:	baobinhduong.vn

At the bottom, there are four buttons: "Delete" (red), "Save and add another" (blue), "Save and continue editing" (blue), and "SAVE" (blue).

Hình 4.21 Giao diện thêm/xóa/sửa một bài viết

4.3.5 Các chức năng của ứng dụng cho người dùng

- Chức năng tìm kiếm bài viết bằng khóa: khi người dùng nhập từ khóa tìm kiếm vào ô tìm kiếm, hệ thống sẽ lọc ra các bài viết liên quan tới từ khóa đó và hiển thị kết quả tìm kiếm lên bảng tin



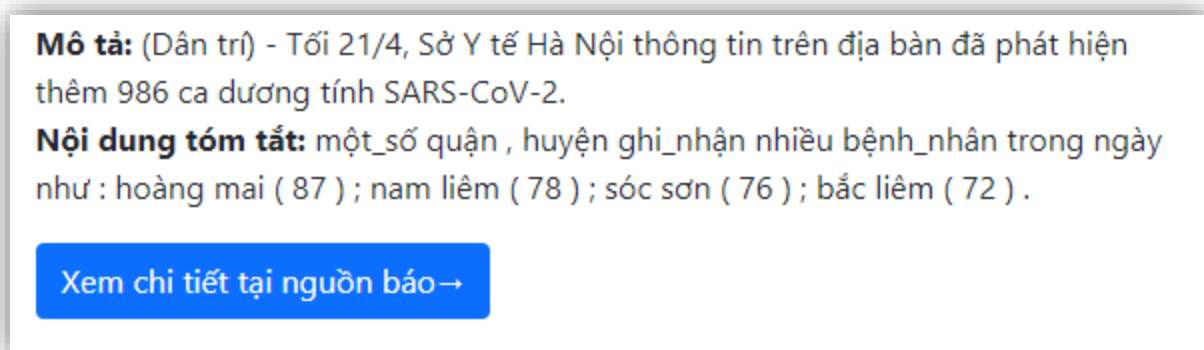
Hình 4.22 Giao diện chức năng tìm kiếm bằng từ khóa

- Chức năng lọc các bài viết theo mặt báo: người dùng chọn mặt báo mong muốn, hệ thống sẽ hiển thị các bài viết tại mặt báo đó



Hình 4.23 Giao diện chức năng lọc bài viết theo mặt báo

- Tương tự như việc lọc bài viết theo mặt báo, khi người dùng chọn các từ khóa nổi bật, hệ thống sẽ tìm kiếm và hiển thị bài viết liên quan tới các từ khóa nổi bật.
- Chức năng đi đến nguồn bài báo covid: khi người dùng đọc và cảm thấy muốn xem chi tiết bài báo, người dùng sẽ click vào button để đi đến nguồn bài báo chứa nội dung chi tiết của bài viết.



Hình 4.24 Giao diện chức năng chuyển hướng tới link gốc bài báo

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Trích rút từ khoá từ trang web là một bài toán hay của hệ thống bài toán trích rút từ khoá cho một văn bản. Ở mức cao hơn, nó là một bài toán con trong hệ thống trích xuất thông tin (Information Retrieval). Trong nhiều năm qua, bài toán này đã được đề cập, quan tâm nhiều ở các hội nghị quốc tế và các công ty lớn. Bài toán trích rút từ khoá từ trang web là việc trích rút từ khoá trong văn bản nội dung trang web. Đây cũng là vấn đề khá mới mẻ và được áp dụng trong rất nhiều lĩnh vực khác nhau như: Hỗ trợ tìm kiếm, hỗ trợ gợi ý người dùng, tóm tắt văn bản, sửa lỗi chính tả, . . .

Trong đề tài “*Xây dựng bảng tin rút gọn về dịch bệnh Covid-19 bằng kỹ thuật Text mining*”, người thực hiện đề tài đã nghiên cứu và áp dụng được những kiến thức đã học về thu thập, xử lý ngôn ngữ tự nhiên bằng ngôn ngữ python, xây dựng ứng dụng website bằng HTML kết hợp với thư viện Bootstrap và Framework Django. Từ đó xây dựng được một bảng tin thu gọn về dịch bệnh Covid, cung cấp một nội dung tóm tắt về bài viết, giúp người dùng có cái nhìn tổng quan về bài báo mà họ đang đọc.

5.2 Hướng phát triển

Đây là một đề tài khá hữu dụng trong thời kỳ công nghệ mới. Dữ liệu được truyền tải trên mạng thông qua các văn bản hàng ngày, hàng giờ, chính vì vậy, hướng phát triển tiếp theo của đề tài có thể là áp dụng với các website tin tức trong nhiều lĩnh vực khác nhau, không chỉ riêng về dịch bệnh Covid-19, thêm các chức năng cơ bản và nâng cao khác như: hệ thống sẽ sử dụng giọng nói AI để đọc cho người dùng nghe nội dung bài viết, người dùng có thể chọn và tìm kiếm nhiều lĩnh vực khác nhau, hoặc hệ thống có thể giúp người dùng tóm tắt một văn bản bất kỳ nào do người dùng nhập vào, . . có thể triển khai ứng dụng trên mạng internet để mọi người có thể truy cập và sử dụng.

PHẦN C: PHỤ LỤC VÀ TÀI LIỆU THAM KHẢO

TÀI LIỆU THAM KHẢO

- [1] Dục Đoàn Trình, Text mining- khai phá dữ liệu từ văn bản, : 13 tháng 1, 2022
- [2] Đặng Phương Mai, Hướng dẫn xử lý ngôn ngữ tự nhiên, ngày 22 tháng 11 năm 2017
- [3] Học viện đào tạo CNTT NIIT - ICT Hà Nộiim Lập trình Web với Django, tháng 09 năm 2020
- [4] Nguyễn Thị Thủy, Tóm tắt văn bản trong Machine Learning, ngày 4 tháng 10 năm 2019
- [5] Prateek Joshi, An Introduction to Text Summarization using the TextRank Algorithm, tháng 11 năm 2018
- [6] Quoc Dinh Truong, Một giải pháp tóm tắt văn bản tiếng Việt tự động, tháng 12 năm 2012
- [7] Trần Ngọc Minh, Machine Learning: Trích xuất dữ liệu từ các trang web dùng thư viện BeautifulSoup, tháng 8 năm 2019