

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Lê Đình Trung - Nguyễn Minh Trường

ỨNG DỤNG
TỔNG HỢP THÔNG TIN ĐỊA PHƯƠNG
TRÊN BÁO CHÍ

KHÓA LUẬN TỐT NGHIỆP CỦ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

Thành phố Hồ Chí Minh, tháng 03/2021

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Lê Đình Trung - 1612751
Nguyễn Minh Trường - 1612760

ỨNG DỤNG
TỔNG HỢP THÔNG TIN ĐỊA PHƯƠNG
TRÊN BÁO CHÍ

KHÓA LUẬN TỐT NGHIỆP CỦ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

GIÁO VIÊN HƯỚNG DẪN
Th.S Văn Chí Nam

Thành phố Hồ Chí Minh, tháng 03/2021

Lời cảm ơn

Chúng tôi xin chân thành cảm ơn Khoa Công nghệ thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh đã tạo điều kiện cho chúng tôi thực hiện khóa luận này.

Đặc biệt, chúng tôi xin bày tỏ lòng biết ơn sâu sắc đến **Th.S Văn Chí Nam** đã tận tình hướng dẫn, thầy đã chỉ bảo cho chúng tôi trong suốt thời gian thực hiện khóa luận này. Chúng tôi đã học được nhiều điều vô giá không chỉ là kiến thức mà còn là cách làm việc, kinh nghiệm sống từ thầy, chắc chắn sẽ giúp ích rất nhiều cho con đường trưởng thành sau này.

Chúng tôi cũng xin gửi lời cảm ơn sâu sắc đến các thầy cô trong Khoa Công Nghệ Thông Tin đã tận tình giảng dạy, trang bị cho chúng tôi những kiến thức quý báu và tạo điều kiện thực hành tốt nhất trong suốt quá trình học tập và nghiên cứu.

Mặc dù chúng tôi đã cố gắng hoàn thành khóa luận trong phạm vi và khả năng cho phép, nhưng chắc chắn sẽ không tránh khỏi những thiếu sót, kính mong nhận được sự bổ sung, góp ý kiến của các thầy giáo, cô giáo và các bạn đề tài khóa luận của chúng tôi được hoàn thiện hơn.

Hồ Chí Minh, ngày 02 tháng 03 năm 2021

Nhóm thực hiện

Lê Đình Trung, Nguyễn Minh Trường

Mục lục

Lời cảm ơn	i
Đề cương chi tiết	ii
Mục lục	ii
Tóm tắt	vi
1 Giới thiệu	1
1.1 Giới thiệu đề tài	1
1.2 Khảo sát	2
1.2.1 Báo Mới	2
1.2.2 Google Tin tức	5
1.2.3 Microsoft Tin tức	9
1.3 Lý do lựa chọn đề tài	9
1.4 Mục tiêu của đề tài	10
1.5 Đề xuất giải pháp	11
1.6 Phạm vi đề tài	12
2 Cơ sở lý thuyết	13
2.1 Tìm hiểu về phương pháp xử lý ngôn ngữ tự nhiên	13
2.1.1 Dánh giá độ tương đồng văn bản	13
2.1.2 Gom nhóm văn bản dựa theo sự tương đồng của chúng	16
2.2 Tách từ và phân lớp văn bản	18

2.2.1	Tiếp cận dựa trên từ điển	19
2.2.2	Tiếp cận dựa trên thống kê	19
2.2.3	Tiếp cận theo hướng kết hợp	19
2.3	Giới thiệu về Node.js	20
2.4	Ngôn ngữ TypeScript	21
2.5	Thư viện giao diện React.js	22
2.6	Giới thiệu về hệ quản trị cơ sở dữ liệu MongoDB	23
2.6.1	Cơ sở dữ liệu NoSQL	23
2.6.2	MongoDB	23
2.7	Giới thiệu về GitHub và GitHub Actions	24
2.7.1	Github	24
2.7.2	GitHub Actions	24
2.8	Hệ thống tổng hợp thông tin trên báo chí	26
2.8.1	Hệ thống tổng hợp tin tức dựa trên công nghệ RSS Feeds	26
2.8.2	Hệ thống tổng hợp tin tức bằng cách cào dữ liệu các trang tin tức	28
2.8.3	Sử dụng trình thu thập dữ liệu web để thu thập tin tức	30
3	Thiết kế	31
3.1	Giải pháp tổng quát	31
3.2	Thiết kế hệ thống	31
3.3	Tổng hợp tin tức	34
3.3.1	Thu thập dữ liệu web từ các URL đầu vào	34
3.3.2	Trình trích xuất tin tức	35
3.3.3	Trích xuất dựa trên giao thức Open Graph	35
3.3.4	Trích xuất sử dụng thư viện Readability	36
3.3.5	Xây dựng trình trích xuất sẵn cho các nguồn tin tức phổ biến	37
3.3.6	Lưu trữ dữ liệu	37

3.4	Lọc tin tức địa phương	38
3.5	Tin chính	38
3.6	Xu hướng tin tức	39
3.7	Bản đồ tin tức	39
3.8	Quản trị thống kê	40
4	Cài đặt	41
4.1	Môi trường thực nghiệm	41
4.2	Mô đun tổng hợp và phân tích tin tức	42
4.3	Dịch vụ RESTful API	43
4.4	Giao diện Front-end	44
4.4.1	Tin chính	44
4.4.2	Xu hướng	45
4.4.3	Bản đồ tin tức	45
5	Kết luận	47
5.1	Kết quả đạt được	47
5.2	Hạn chế	47
5.3	Hướng phát triển trong tương lai	48
Tài liệu tham khảo		50

Danh sách hình

1.1	Giao diện trang chủ Báo mới.	3
1.2	Giao diện trên thiết bị di động của chức năng Tin địa phương	4
1.3	Giao diện trang chủ Google Tin tức Việt Nam	6
1.4	Giao diện chuyên mục Thông tin toàn cảnh của Google Tin tức	7
1.5	Giao diện mục Tin tức địa phương của Google Tin tức . .	8
2.1	Node.js là nền tảng lập trình phổ biến nhất năm 2020, theo Stack Overflow 2020 Developer Survey [7]	21
2.2	Biểu tượng nhận diện một trang web có hỗ trợ RSS	27
2.3	Mô hình một trình thu thập dữ liệu web đơn giản	30
3.1	Sơ đồ tình huống sử dụng	32
3.2	Sơ đồ tình huống sử dụng	33
3.3	Thu thập và trích xuất tin tức	34
3.4	Lưu trữ dữ liệu	37
4.1	GitHub Actions thực thi tác vụ theo chu kỳ	44
4.2	Giao diện trang chủ Tin chính	45
4.3	Giao diện Xu hướng	46
4.4	Giao diện Bản đồ tin tức	46

Danh sách bảng

1.1	Chi tiết các chức năng	12
4.1	Các thư viện được sử dụng	42

Tóm tắt khóa luận

Khóa luận trình bày những kiến thức và kỹ thuật mà chúng tôi đã tìm hiểu để thực hiện đề tài "Ứng dụng tổng hợp thông tin địa phương trên báo chí", cũng như cách thức triển khai ứng dụng minh họa. Khóa luận được chia thành 5 chương, bao gồm:

Chương 1 - Giới thiệu: Trình bày ngữ cảnh và vấn đề phải giải quyết, lý do chọn đề tài, mục tiêu và phạm vi phát triển của đề tài, khảo sát các sản phẩm có trên thị trường, phát hiện những ưu điểm và nhược điểm của chúng, từ đó đưa ra giải pháp thay thế.

Chương 2 - Cơ sở lý thuyết: Trình bày lý thuyết, khái niệm liên quan về "Hệ thống tổng hợp thông tin trên báo chí" và "Phát hiện tin tức địa phương".

Chương 3 - Thiết kế: Trình bày cơ sở giải pháp kỹ thuật, công nghệ có sử dụng trong hệ thống để phát triển các chức năng của ứng dụng.

Chương 4 - Cài đặt: Trình bày chi tiết cách thức xây dựng, cài đặt các công cụ để hiện thực từng giải pháp đã nêu trong chương 3.

Chương 5 - Kết luận: Liệt kê và đánh giá kết quả đã thu được sau khi thực hiện khóa luận, cũng như quá trình xây dựng hệ thống. Ngoài ra, chương này cũng trình bày phương hướng phát triển của hệ thống trong tương lai.

Chương 1

Giới thiệu

1.1 Giới thiệu đề tài

Trong thời đại công nghệ thông tin, những trang báo điện tử đã trở nên phổ biến và quen thuộc với độc giả. Ngày nay, chỉ cần một chiếc điện thoại thông minh là độc giả có thể đọc được hàng trăm tin tức nóng khác nhau mỗi ngày. Theo thống kê, tính đến ngày 31/12/2020, Việt Nam hiện có 779 cơ quan báo chí, bao gồm 142 báo, 612 tạp chí và 25 cơ quan báo chí điện tử độc lập [1].

Số lượng lớn là vậy nhưng các trang báo đa số đều không có chức năng xem tin theo khu vực địa lý. Điều này khiến người dùng tại một địa phương mong muốn đọc các tin tức liên quan tới địa phương của mình thì phải tìm kiếm rất khó khăn và thông tin thường không đầy đủ.

Bên cạnh đó, xuất phát từ nhu cầu của cơ quan quản lý địa phương cần thu thập thông tin liên quan đến địa phương mình để nắm bắt, quản lý kịp thời. Việc theo dõi thường xuyên và nắm bắt thông tin nhanh nhất từ các kênh tin tức là vấn đề cấp thiết để các cơ quan chức năng nắm bắt tình hình địa phương.

Đề tài hướng đến việc xây dựng của một hệ thống có khả năng thu thập nội dung dữ liệu từ mạng Internet, thu thập bài viết từ các trang báo, cung cấp thông tin điện tử. Hệ thống cung cấp giao diện trang web để

người dùng truy cập và đọc những tin tức địa phương đã được tổng hợp đó. Ngoài ra, để đảm bảo chất lượng thông tin tới độc giả, hệ thống cung cấp các chức năng lọc nội dung và phân loại các bài viết theo từng khu vực, địa phương, thống kê tin tức theo các từ khóa được quan tâm.

1.2 Khảo sát

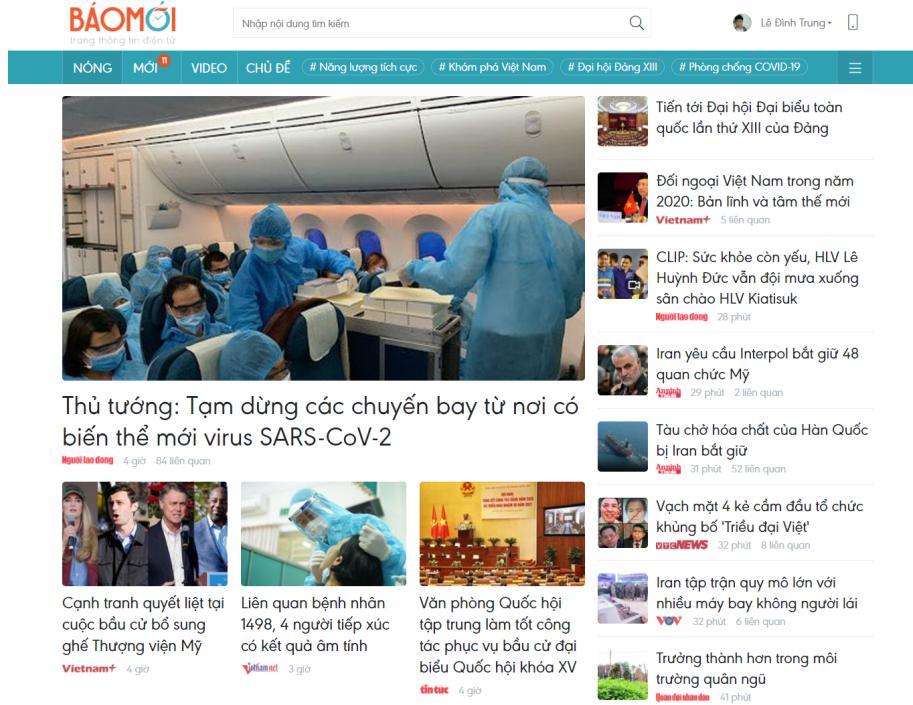
Hiện nay có một số bên cung cấp giải pháp tương tự, sau đây chúng tôi xin liệt kê và phân tích điểm mạnh, yếu của từng giải pháp, bao gồm trang tin Báo Mới, Google Tin tức và Microsoft Tin tức.

1.2.1 Báo Mới

Báo Mới là một trang mạng tổng hợp tin tức Tiếng Việt được điều khiển tự động bởi máy tính phát triển bởi EPI*. Báo Mới hiện có thể truy cập tại địa chỉ <https://baomoi.com>. Với gần 200 nguồn chính thức từ các báo điện tử và trang tin điện tử Việt Nam được Báo Mới tự động tổng hợp, phân loại, phát hiện các bài đăng lại, nhóm các bài viết liên quan và hiển thị theo sở thích đọc tin của từng độc giả. Theo thống kê từ Alexa, Báo Mới nằm trong top 30 các trang web được truy cập nhiều nhất Việt Nam năm 2020 [2].

Các tính năng chính:

- Phân loại nội dung: Hệ thống tự động phân tích nội dung các tin tức và phân loại vào chuyên mục thích hợp.
- Phát hiện bài trùng lặp: Hệ thống tự động phát hiện các bài đăng lại và nhóm chúng lại về bài nội dung gốc.
- Nhóm các bài liên quan: Hệ thống tự động phát hiện các bài liên quan (không phải bài đăng lại) về cùng một chủ đề nào đó.



Hình 1.1: Giao diện trang chủ Báo mới.

- Bóc tách từ khóa: Hệ thống tự động tách ra các từ khóa (keyword) của bài viết, giúp người đọc dễ dàng tìm kiếm các thông tin liên quan đa chiều.

Tính năng tin địa phương

Báo Mới có tính năng đọc tin địa phương khi người dùng truy cập bằng thiết bị di động và thực hiện chọn mục tin địa phương. Sau khi người dùng nhấn chọn thì giao diện hiển thị danh sách các tỉnh thành để người dùng lựa chọn, với mỗi lựa chọn thì Báo Mới sẽ hiển thị các tin địa phương của tỉnh, thành đó. Tuy nhiên, chức năng này mới chỉ dừng lại ở cấp độ tỉnh thành phố, khá rộng lớn đối với nhu cầu của người dân hiện nay (mức độ quận huyện).

Báo Mới có ưu điểm là lượng nguồn tin phong phú và cập nhật thường xuyên. Báo Mới còn trích xuất thông tin bài để cho ra các từ khoá, từ các từ khoá đó, người dùng có thể xem các tin bài liên quan. Tuy nhiên, chúng tôi nhận thấy trang báo này cũng tồn tại các yếu điểm sau:

TP. HCM



Khánh thành 'Không gian xanh' của sinh viên Ký túc xá ĐHQG TP. HCM

Sinhviên 3 giờ



TP HCM: Truy vết, cách ly 4 người tiếp xúc du học sinh mắc Covid-19

Nguoi lao dong 3 giờ



Trường ĐH Kinh tế TP. HCM tuyển 6.350 chỉ tiêu ở năm 2021

Sinhviên 6 giờ



CLIP: Một số hình ảnh về vụ nổ ở TP HCM do Tổ chức khủng bố 'Triều đại Việt' gây ra

Nguoi lao dong 2 giờ

Hình 1.2: Giao diện trên thiết bị di động của chức năng Tin địa phương

- Tin tức địa phương chỉ hỗ trợ trên thiết bị di động, không hỗ trợ xem trên máy tính.
- Tin tức địa phương trên khu vực quá lớn (Tỉnh hoặc thành phố).
- Tin tức địa phương không có các tính năng thống kê riêng biệt, người dùng không thể biết xu hướng tin tức tại địa phương mình.
- Các tính năng chính của hệ thống không hoạt động trong chế độ xem tin địa phương.

1.2.2 Google Tin tức

Google Tin tức (Google News) là một trang mạng tổng hợp tin tức tự động được cung cấp bởi Google. Google Tin tức được phát hành toàn cầu, tại Việt Nam, Google Tin tức có sẵn tại địa chỉ <https://news.google.com.vn>

Tính năng xem thông tin toàn cảnh

Một tính năng nổi bật của trang web này là tính năng xem thông tin toàn cảnh. Các bài viết cùng chủ đề được hệ thống nhận diện và gom nhóm lại, người dùng có thể đọc chúng trong một trang chung, được gọi là Trang thông tin toàn cảnh. Tính năng này cho phép người dùng đọc những tin tức liên quan một cách nhanh chóng, cung cấp cái nhìn tổng quan về một chủ đề nhất định.

Tính năng tin địa phương

Google Tin tức cũng cung cấp tính năng tin địa phương khi người dùng chọn mục Tin địa phương. Hiện tính năng này chỉ hỗ trợ các khu vực là thành phố lớn.

Google Tin tức có điểm mạnh là hoạt động nhanh, giao diện hiện đại. Việc gom tin tức thành các nhóm chủ đề một cách trực quan, giúp người

≡ Google Tin tức

Tìm kiếm chủ đề, vị trí và nguồn

Tin chính

Bầu cử Mỹ: Đảng Cộng hòa nội chiến

Người Lao Động · 3 giờ trước

- Tổng thống Trump, ông Biden 'quyết đấu' lần cuối giành kiểm soát Thượng viện tại Georgia
- Báo Thanh Niên · 11 giờ trước
- Biden đứng trước 'ngưỡng cửa lịch sử' tại Georgia
- VnExpress · 17 giờ trước
- Chuyên gia Trung Quốc mổ xẻ quan hệ Mỹ - Trung thời Biden
- VietNamNet · 19 giờ trước
- Chuyên cơ 'Không lực Hai' của Tổng thống Trump đến Scotland ngay trước ngày ông Biden nhậm chức?
- Báo Thanh Niên · 11 giờ trước

Xem Thông tin toàn cảnh

Iran nói bắt tàu dầu có 2 thủy thủ Việt không phải để trả đũa Hàn Quốc

Báo Thanh Niên · 4 giờ trước

- Đòn Iran thả tàu, Hàn Quốc điều gấp tàu chiến áp sát eo biển Hormuz
- Tuổi Trẻ Online · 19 giờ trước

Xem Thông tin toàn cảnh

Tin chính khác

Hồ Chí Minh

Có mây rải rác

25°c

Hôm nay	Th 5	Th 6	Th 7	CN
33°C 23°C	33°C 23°C	34°C 22°C	33°C 21°C	33°C 21°C

c | f | k weather.com

Thời sự

Phạm Chí Dũng Donald Trump

I-ran Việt Nam

Tổng thống Hoa Kỳ Joe Biden

Bộ Công an Phi-lip-pin

Đảng Cộng hòa

Hội Nhà báo Độc lập Việt Nam

Hình 1.3: Giao diện trang chủ Google Tin tức Việt Nam

Thông tin toàn cảnh

Sắp xếp ▾

Chia sẻ

Tất cả bài viết

Biển thẻ nCoV ở Nam Phi đáng sợ hơn biển thẻ từ Anh

VnExpress · 1 giờ trước



Thủ tướng: Tạm dừng chuyến bay từ nơi có biến thể mới virus SARS-CoV-2 | VTC Now

VTC NOW · 3 giờ trước



Lịch trình di chuyển của du học sinh mắc Covid-19 ở Hạ Long, BN 1498 | VTC Now

VTC NOW · 4 giờ trước



Quảng Ninh truy vết, cách ly người tiếp xúc với ca nghi mắc Covid-19

Truyền Hình Nhân Dân · 4 giờ trước



Tạm dừng chuyến bay từ các nước có Covid-19 chủng mới về Việt Nam

Dân Trí Mobile · 5 giờ trước



Hình 1.4: Giao diện chuyên mục Thông tin toàn cảnh của Google Tin tức

Tin tức địa phương của bạn

[\(?\) Tại sao lại hiển thị những địa điểm này?](#)



Ho Chi Minh



phường 1



Thêm

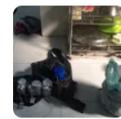
Tết mà ở lại phòng trọ Sài Gòn chắc buồn lắm!

Báo Thanh Niên · 5 giờ trước



CLIP: Một số hình ảnh về vụ nổ ở TP HCM do Tổ chức khủng bố 'Triều đại Việt' gây ra

Người Lao Động · 3 giờ trước



TP.HCM: 45 dự án giao thông trọng điểm hoàn thành năm 2021

Báo Pháp Luật TP.HCM · 8 giờ trước



Phó chủ tịch UBND TP HCM Lê Hòa Bình trao quyết định cho ông Lê Ngọc Hùng

Người Lao Động · 10 giờ trước



Hình 1.5: Giao diện mục Tin tức địa phương của Google Tin tức

dùng có cái nhìn tổng quan về xu hướng tin tức. Tuy nhiên, ứng dụng này cũng có các điểm yếu:

- Tin tức trên trang chủ cập nhật và phản ứng chậm: Nội dung các tin chính trên trang chủ hầu ít thay đổi trong ngày.
- Tin tức địa phương trên khu vực lớn (Tỉnh, thành phố lớn).
- Tin tức địa phương không có các tính năng thống kê riêng biệt, người dùng không thể biết xu hướng tin tức tại địa phương mình.
- Tin tức địa phương đôi khi chưa chính xác.
- Người dùng không thể tùy chỉnh bất cứ điều gì liên quan tới thuật toán cung cấp tin tức.

1.2.3 Microsoft Tin tức

Microsoft Tin tức (Microsoft News), tiền thân là MSN Tin tức, được phát triển bởi Microsoft. Ra mắt vào ngày 26 tháng 10 năm 2012, Microsoft Tin tức là một công cụ tổng hợp tin tức và dịch vụ có các tiêu đề và bài viết được lựa chọn bởi các biên tập viên.

Microsoft Tin tức đưa tin dựa theo công nghệ RSS cổ điển, điều này khiến thông tin cập nhật chậm và không đầy đủ. Ngoài ra ứng dụng này cũng không hỗ trợ đưa tin địa phương và xem các tin tức theo chủ đề. Với khá nhiều nhược điểm như vậy, Microsoft khá thua thiệt khi so sánh với 2 sản phẩm kể trên là Google Tin tức và Báo mới.

1.3 Lý do lựa chọn đề tài

Qua khảo sát các sản phẩm có sẵn trên thị trường, chúng tôi nhận ra đề tài "Ứng dụng tổng hợp thông tin địa phương trên báo chí" là một đề tài mới mẻ và triển vọng bởi vì các hệ thống tổng hợp tin tức này đều đưa

tin dàn trải, không chú trọng vào một địa phương cụ thể. Do đó, thông tin về một địa phương bị phân tán, người dùng muốn tìm các tin tức về địa phương mình phải thực hiện tìm kiếm với độ chính xác thấp. Ứng dụng đánh mạnh vào nhu cầu đó của người dùng, tập trung phục vụ việc đưa tin về một địa phương nhất định. Bên cạnh đó ứng dụng còn có các chức năng nổi bật sau:

- Chức năng quản trị cho nhà quản lý địa phương: Nhà quản lý có thể xem thống kê về xu hướng tin tức, loại bỏ một thông tin hoặc nguồn thông tin không chính xác.
- Chức năng Bản đồ thông tin: Ứng dụng cung cấp giao diện đồ họa bản đồ trực quan về các tin tức theo các khu vực nhỏ hơn. Từ đó người dùng có cái nhìn tổng quát về những gì đang xảy ra trên địa phương mình lúc này.
- Phát hiện xu hướng: Thể hiện các chủ đề nóng đang xảy ra trên khu vực.

1.4 Mục tiêu của đề tài

Xây dựng ứng dụng tổng hợp thông tin báo chí theo địa phương. Người dùng thuộc địa phương có thể đọc tin tức về địa phương của mình, tùy chọn tin theo các chủ đề yêu thích. Ứng dụng cũng cung cấp chức năng quản lý dành riêng cho chính quyền địa phương để nắm bắt chính xác thông tin của địa phương do mình quản lý.

Ứng dụng có các chức năng chính:

- Tự động cập nhật tin tức mới từ các trang báo chí phổ biến.
- Lọc tin theo địa phương.
- Lọc theo các thẻ, từ khóa.

- Thông tin cung cấp từ nhiều nguồn tin chính thống.

Các chức năng mở rộng:

- Phát hiện các chủ đề tin tức nóng.
- Bản đồ tin tức thể hiện trực quan độ nóng theo khu vực (dựa theo số lượng tin tức tại khu vực đó).
- Tùy biến tin tức theo thói quen người dùng.
- Phân tích tin tức: đánh giá tích cực hay tiêu cực
- Chức năng quản trị và thống kê dành cho chính quyền địa phương.

1.5 Đề xuất giải pháp

Dựa theo kết quả phân tích các sản phẩm tương tự trên thị trường hiện nay, chúng tôi nhận định đề tài có khả năng thành công lớn khi tập trung vào chức năng tin tức địa phương, một chức năng mà các sản phẩm khác không có hoặc không đầy mạnh. Để nhấn mạnh hơn về sự khác biệt so với các sản phẩm khác, chúng tôi quyết định đặt tên, cũng như tên miền của ứng dụng là "Tin địa phương", có địa chỉ tại <https://tindiaphuong.github.io>

Sau đây là một số chức năng sẽ được hiện thực trong ứng dụng:

STT	Tên chức năng	Mô tả chức năng
1	Xem tin địa phương hàng đầu	Trên trang chủ, hiển thị danh sách các tin địa phương hàng đầu.
2	Gom tin theo nhóm	Các tin tức cùng một chủ đề được gom lại chung với nhau trong cùng một mục.
3	Xác thực người dùng	Người dùng có thể đăng ký tài khoản để sử dụng các chức năng nâng cao.

4	Tài khoản quản trị viên	Tài khoản quản trị viên cho phép nhà quản lý địa phương xem thống kê tin tức, quản lý hay kiểm duyệt tin tức.
5	Bản đồ tin tức	Hiển thị trực quan xu hướng tin tức theo mỗi khu vực nhỏ trên một bản đồ nhỏ
6	Xu hướng tin tức	Hiển thị các chủ đề nóng nhất trên một đồ họa trực quan
7	Tìm kiếm	Tìm kiếm thông tin liên quan bằng cách nhập vào một từ khóa

Bảng 1.1: Chi tiết các chức năng

1.6 Phạm vi đề tài

- Ứng dụng chỉ tổng hợp thông tin một địa phương (quận, huyện).
Ứng dụng không có chức năng tổng hợp thông tin ngoài địa phương đã xác định đó.
- Ứng dụng web có thể chạy trên trình duyệt web Google Chrome, Mozilla Firefox, Microsoft Edge và Safari.

Chương 2

Cơ sở lý thuyết

2.1 Tìm hiểu về phương pháp xử lý ngôn ngữ tự nhiên

2.1.1 Đánh giá độ tương đồng văn bản

Khi chúng tôi thực hiện hiển thị các tin tức mà trình tổng hợp tin tức thu thập được lên giao diện đồ họa, chúng tôi nhận thấy rằng, trong một thời điểm, có rất nhiều tin tức, từ nhiều nguồn tin khác nhau, có sự liên hệ và tương đồng với nhau. Các tin tức này khi hiển thị đồng thời cùng lúc khiến người dùng bị "choáng ngợp", mang lại trải nghiệm không tốt. Điều này dẫn tới nhu cầu cần thiết rằng, các tin như vậy cần được xác định và được gom thành một nhóm. Hệ thống từ đó chỉ hiển thị một hoặc một số tin tức nổi bật, đồng thời hệ thống cũng cung cấp công cụ để người dùng đọc được toàn bộ các tin được gom nhóm liên quan tới tin tức đó.

Để hiện thực điều này, chúng tôi cần đánh giá, một cách tương đối sự tương đồng giữa 2 văn bản với nhau, từ đó tiến hành gộp chung lại thành một nhóm chung. Đây là bài toán đánh giá độ tương đồng giữa hai mẫu văn bản trong xử lý ngôn ngữ tự nhiên. Hiện nay có nhiều thuật toán có sẵn để giải quyết bài toán này.

Thuật toán khoảng cách Levenshtein

Khoảng cách Levenshtein [3] thể hiện khoảng cách khác biệt giữa hai chuỗi ký tự. Khoảng cách này được đặt theo tên người đề ra khái niệm này, Vladimir Levenshtein. Nó được sử dụng trong việc tính toán sự giống nhau và khác nhau giữa hai chuỗi, ví dụ như chương trình kiểm tra và gợi ý sửa lỗi chính tả.

Khoảng cách Levenshtein giữa chuỗi a và chuỗi b là số bước ít nhất để biến đổi chuỗi a thành chuỗi b thông qua ba phép biến đổi:

- Xóa một ký tự.
- Thêm một ký tự.
- Thay ký tự này bằng ký tự khác.

Khoảng cách giữa hai chuỗi "kitten" và "sitting" là 3 vì chúng ta cần ít nhất 3 lần biến đổi:

- Thay "k" bằng "s"
- Thay "e" bằng "i"
- Thêm ký tự "g" vào cuối.

Ta sử dụng thuật toán quy hoạch động để tính toán Khoảng cách Levenshtein. Sau đây là mã giả giải thuật Wagner-Fischer, tính toán trên mảng hai chiều $(n + 1) \times (m + 1)$ với n, m là độ dài của 2 chuỗi a, b cần tính khoảng cách.

Thuật toán khoảng cách Jaro - Winkler

Khoảng cách Jaro - Winkler [4] được phát triển bởi William E.Winkler (1990), cải tiến từ khoảng cách Jaro (1989) là một thuật toán đo lường khoảng cách giữa 2 mẫu văn bản. Độ tương đồng Jaro - Winkler là giá trị

Algorithm 1 Khoảng cách Levenshtein

```
1: function LEVDIST( a, b)
2:   n  $\leftarrow$  a.size();
3:   m  $\leftarrow$  b.size();
4:   distance[i, 0]  $\leftarrow$  i  $\forall$  i  $\in$  0..n;
5:   distance[0, j]  $\leftarrow$  j  $\forall$  j  $\in$  0..m;
6:   for j  $\leftarrow$  1 to m do
7:     for i  $\leftarrow$  1 to n do
8:       indicator  $\leftarrow$  a[i]  $\neq$  b[j] ? 1 : 0;
9:       distance[i, j]  $\leftarrow$  Minimum(
10:        distance[i - 1, j] + 1,
11:        distance[i - 1, j] + 1,
12:        distance[i - 1, j] + 1 + indicator);
13:   return distance[n; m];
```

ngược lại với khoảng cách Jaro - Winkler, thể hiện độ tương đồng giữa hai mẫu văn bản. Để tính giá trị này, trước hết cần tính độ tương đồng Jaro:

$$sim_j = \begin{cases} 0 & \text{nếu } m = 0 \\ \frac{1}{3} \left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m} \right) & m \text{ khác } 0 \end{cases}$$

Với:

- *m* là số lượng ký tự (trùng khớp). Hai ký tự từ *a* và *b* được gọi là **trùng khớp** nếu chúng giống nhau và vị trí của chúng trong chuỗi không cách nhau xa quá $\left\lfloor \frac{\max(|a|, |b|)}{2} \right\rfloor - 1$ ký tự
- *t* là một phần hai số lượng ký tự "trùng khớp" nhưng khác vị trí.

Từ giá trị độ tương đồng Jaro, ta có thể tính toán độ tương đồng Jaro - Winkler được định nghĩa bởi:

$$sim_{jw}(a, b) = sim_j(a, b) + l * p * (1 - sim_j(a, b))$$

Với:

- l là số ký tự tính từ đầu tới ký tự đầu tiên giống nhau của cả hai chuỗi, tối đa là 4.
- p là hằng số biến dạng. Giá trị này thông thường là 0.1 và có thể điều chỉnh để thay đổi biên độ giá trị trả về, nhưng không quá 0.25

Độ tương đồng Jaro - Winkler có miền giá trị từ 0 tới 1. Nếu hai chuỗi giống nhau hoàn toàn, giá trị của nó là 1. Ngược lại, nếu hai chuỗi khác nhau hoàn toàn, giá trị của nó là 0.

Hệ số Sørensen–Dice

Hệ số Sørensen–Dice, hay chỉ số Sørensen–Dice là một giá trị được dùng trong hoạt động thống kê để đo sự tương đồng giữa hai mẫu, được phát triển độc lập nhau bởi Thorval Sørensen và Lee Raymond Dice vào năm 1948 và 1945.

Độ tương đồng giữa hai chuỗi a và b được tính bởi:

$$d = \frac{2n_t}{(n_a + n_b)}$$

Với:

- n_t là số **bigram** tìm thấy trong cả hai chuỗi. **Bigram** là tên gọi chỉ cụm 2 ký tự liền nhau và thường xuyên xuất hiện trong một chuỗi. Ví dụ trong tiếng anh, các bigram là "th", "he", "in", "an" và "er" là thường hay xuất hiện nhất trong câu.
- Tương tự, n_a , n_b là số **bigram** tìm thấy trong chuỗi a và b .

2.1.2 Gom nhóm văn bản dựa theo sự tương đồng của chúng

Phân tích cụm hay phân nhóm, gom cụm (Cluster analysis, Clustering) [5] là một tác vụ gom nhóm một tập các đối tượng theo cách các đối

tượng cùng nhóm (cluster) có tính giống nhau theo một đặc tính nào đó hơn so với các đối tượng ngoài nhóm hoặc thuộc các nhóm khác.

Bài toán gom nhóm các văn bản theo sự tương đồng của chúng, chính là một bài toán phân tích cụm, với mỗi cụm có sự tương đồng về nội dung văn bản. Hiện nay có nhiều giải thuật để giải quyết bài toán này, chúng tôi sau đây xin giới thiệu một số giải thuật tiêu biểu và thông dụng, mà chúng tôi đã thử nghiệm thử nghiệm và áp dụng vào khóa luận này:

K-Means

Thuật toán phân hoạch K-Means [6] do MacQueen đề xuất trong lĩnh vực thống kê năm 1967, mục đích của thuật toán K-Means là sinh ra k cụm dữ liệu C_1, C_2, \dots, C_k từ một tập dữ liệu chứa n đối tượng trong không gian d chiều.

Tóm tắt giải thuật:

Đầu vào: Dữ liệu X và số lượng cụm cần tìm k .

Đầu ra: Các tâm M và các nhãn véc tơ cho từng điểm dữ liệu Y .

1. Chọn k điểm bất kỳ làm các tâm ban đầu.
2. Phân mỗi điểm dữ liệu vào cụm có tâm gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cụm ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật tâm cho từng cụm bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cụm đó sau bước 2.
5. Quay lại bước 2.

DBSCAN

Giải thuật DBSCAN (Density Based Spatial Clustering of Application with Noise) được Ester, Kriegel và Sander đề xuất năm 1996 khi nghiên cứu các thuật toán gom cụm dữ liệu. Giải thuật DBSCAN phát hiện các cụm có hình dạng tùy ý, khả năng phát hiện nhiễu tốt. DBSCAN thích hợp với cơ sở dữ liệu có mật độ phân bố dày đặc kể cả có phần tử nhiễu. DBSCAN là thuật toán phân cụm dựa trên mật độ thông dụng nhất, thuật toán đi tìm các đối tượng mà có số đối tượng láng giềng lớn hơn một ngưỡng tối thiểu. Tìm tất cả các đối tượng mà các láng giềng của nó thuộc về lớp các đối tượng đã xác định ở trên, một cụm được xác định bằng một tập tất cả các đối tượng liên thông mật độ với các láng giềng của nó.

2.2 Tách từ và phân lớp văn bản

Tách từ là một phần quan trọng của tính năng **Xu hướng**. Tính năng **Xu hướng** xếp hạng các từ khóa dựa theo tần suất xuất hiện của chúng trong các tin bài tại một thời điểm, từ đó xếp hạng các tin bài có chứa các từ khóa đó. **Xu hướng** yêu cầu hệ thống tách nội dung tin tức thành các từ, cụm từ riêng biệt, sau đó phát hiện, lưu lại các từ khóa quan trọng.

Tách từ là một bài toán quan trọng trong xử lý ngôn ngữ tự nhiên, nhằm xác định được ranh giới các từ có trong văn bản. Trong tiếng việt, ngoài từ đơn (có một âm tiết), còn có từ ghép (nhiều âm tiết). Điều này gây khó khăn cho việc tách từ tự động một cách chính xác do không thể dùng khoảng trắng để xác định ranh giới của các từ. Những âm tiết được kết hợp để tạo thành các từ khác nhau lại phụ thuộc vào ngữ cảnh của văn bản. Để nhận dạng đúng ranh giới của các từ (tách từ), nhiều phương pháp đã được đề xuất. Các phương pháp có thể chia thành ba nhóm chính: dựa trên từ điển, dựa trên thống kê, và phương pháp kết hợp (hybrid).

2.2.1 Tiếp cận dựa trên từ điển

Dựa trên từ điển có sẵn, phương pháp tách từ thực hiện so khớp từng âm tiết trong văn bản với các từ có trong từ điển. Các cách thức so khớp bao gồm:

- So khớp từ dài nhất.
- So khớp từ ngắn nhất.
- So khớp chồng lấp.
- So khớp cực đại.

Phương pháp tách từ dựa trên từ điển có ưu điểm là nhanh, đơn giản và dễ hiểu. Tuy nhiên, độ chính xác của phương pháp này phụ thuộc vào độ phong phú của từ điển được xây dựng. Trong các tình huống xuất hiện từ mới không tồn tại trong từ điển, phương pháp này khó có thể xử lý được.

2.2.2 Tiếp cận dựa trên thống kê

Xây dựng mô hình ngôn ngữ

Với cách tiếp cận dựa trên thống kê, các giải pháp cho việc tách từ thông thường dựa trên mô hình ngôn ngữ (language model). Một mô hình ngôn ngữ thường được xây dựng trên việc thu thập thống kê số lần xuất hiện hoặc cùng xuất hiện của các từ trong một tập hợp lớp các văn bản.

2.2.3 Tiếp cận theo hướng kết hợp

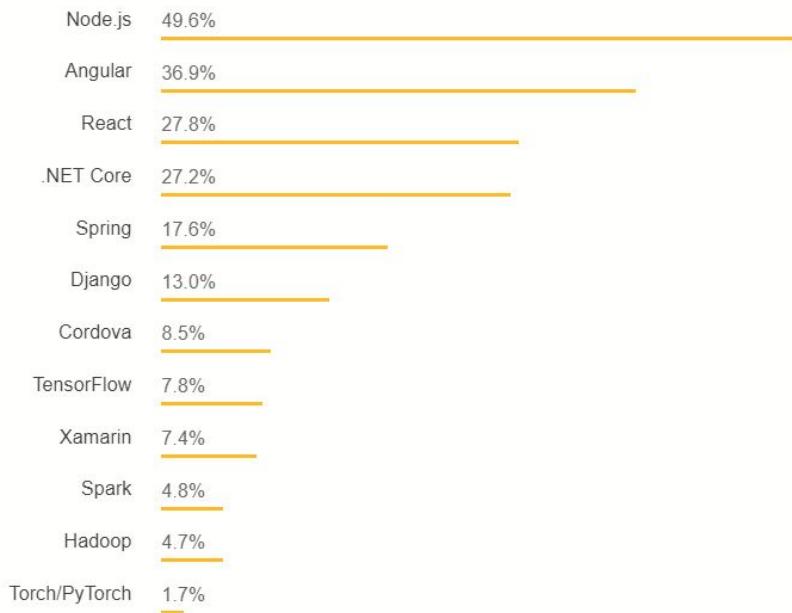
Các phương pháp tiếp cận từ điển và tiếp cận thống kê đều có những ưu điểm, nhược điểm riêng. Do vậy, để tận dụng được những ưu điểm của mỗi loại tiếp cận, các phương pháp kết hợp chúng đã ra đời.

2.3 Giới thiệu về Node.js

Node.js là một hệ thống môi trường phần mềm mã nguồn mở, được thiết kế để viết các ứng dụng chạy trên máy chủ. Một chương trình phần mềm Node.js được viết bằng ngôn ngữ JavaScript, sử dụng kỹ thuật điều khiển theo sự kiện, nhập xuất không đồng bộ để tối thiểu chi phí và tối đa khả năng mở rộng.

Sự phổ biến ngày càng tăng của JavaScript đã kéo theo rất nhiều thay đổi và bộ mặt của sự phát triển Web ngày nay đã khác hẳn. JavaScript xuất hiện ở mọi nơi và sự xuất hiện của Node.js như một cuộc cách mạng về môi trường phát triển đầu tiên hỗ trợ JavaScript trên cả phía máy chủ và trên trình duyệt (Client-Server Side). Ngoài lợi thế sử dụng ngôn ngữ JavaScript, Node.js có các điểm mạnh sau khi so sánh với các nền tảng khác:

- Nhận và xử lý nhiều kết nối đơn luồng (Single-thread): Điều này giúp hệ thống ít tốn Ram và chạy nhanh nhất do không phải khởi tạo luồng mới cho mỗi truy vấn.
- Lợi thế khi xây dựng RESTful API (JSON): Node.js xử lý non-blocking sử dụng JavaScript, cho phép xử lý định dạng JSON một cách nhẹ nhàng và dễ dàng.
- Ứng dụng thời gian thực: Các ứng dụng Node.js đáp ứng tốt thời gian thực và chạy đa nền tảng, đa thiết bị.
- Cộng đồng hỗ trợ lớn: Là một trong những nền tảng phổ biến nhất, Node.js có cộng đồng hỗ trợ lớn, bất cứ vấn đề nào liên quan đều có thể được giải đáp cực kỳ nhanh lẹ.



Hình 2.1: Node.js là nền tảng lập trình phổ biến nhất năm 2020, theo Stack Overflow 2020 Developer Survey [7]

2.4 Ngôn ngữ TypeScript

TypeScript là một ngôn ngữ lập trình được phát triển và duy trì bởi Microsoft. TypeScript với các bổ sung tùy chọn kiểu tĩnh và lớp trên cơ sở lập trình hướng đối tượng. TypeScript được thiết kế để phát triển ứng dụng lớn, được biên dịch hoàn toàn sang JavaScript. Do biên dịch sang JavaScript nên bất kỳ chương trình JavaScript nào cũng đều là chương trình TypeScript hợp lệ. Trong quá trình tìm hiểu về nền tảng Node.js, chúng tôi nhận thấy JavaScript tồn tại các yếu điểm như không hỗ trợ kiểu tĩnh, không hỗ trợ hướng đối tượng, và TypeScript chính là sự thay thế hoàn hảo khi khắc phục các tồn tại của JavaScript, hoàn toàn tương thích với JavaScript, do vậy hoàn toàn tương thích với Node.js

2.5 Thư viện giao diện React.js

React.js hay React là một thư viện mã nguồn mở JavaScript để xây dựng giao diện người dùng (front-end). Thư viện React.js được phát triển bởi Facebook và cộng đồng nhà phát triển.

Các tính năng chính của React.js:

- Viết ứng dụng trực tiếp trên JavaScript.
- Phá vỡ những cấu tạo giao diện phức tạp và chia chúng thành những "Thành phần" độc lập (Component).
- Thay đổi trạng thái cho nhiều Thành phần con trên ứng dụng nhưng không ảnh hưởng tới các Thành phần cha hoặc các thành phần khác.

Chia ứng dụng thành các "Thành phần" lưu trạng thái và có thể tái sử dụng

Trang web sử dụng React.js được chia thành những Thành phần nhỏ (Component). Chúng ta có thể tái sử dụng một Component ở nhiều nơi, với các trạng thái hoặc các thuộc tính khác nhau, trong một Component lại có thể chứa thành phần khác. Mỗi Component trong React.js có một trạng thái riêng, có thể thay đổi, và React.js sẽ thực hiện cập nhật Component dựa trên những thay đổi của trạng thái.

Phản ứng khi có thay đổi

Khi trạng thái của một Thành phần thay đổi, những thay đổi này sẽ được React kiểm soát và tự động cập nhật (react) giao diện mới phù hợp với trạng thái mới.

Cây Dom ảo (VirtualDom)

Việc sử dụng JavaScript để tạo ra mã HTML cho phép React.js có một cây HTML ảo (VirtualDom). Khi một trang web sử dụng React.js được

tải lại, một VirtualDom được tạo ra và lưu trong bộ nhớ. Mỗi khi có một trạng thái thay đổi, một cây đối tượng HTML mới được tạo ra và hiển thị lên trình duyệt. Để tạo lại cây đối tượng HTML mới, thay vì tạo lại toàn bộ cây, React.js dùng một thuật toán thông minh để tạo lại chỉ các thành phần khác biệt giữa cây cũ và cây mới. Và bởi quá trình này xảy ra trong bộ nhớ nên cực kỳ nhanh chóng. Hệ quả trang web không cần phải tải lại hoàn toàn mà chỉ thay đổi trên những "Thành phần" có sự thay đổi, giúp tiết kiệm tài nguyên hệ thống.

2.6 Giới thiệu về hệ quản trị cơ sở dữ liệu MongoDB

2.6.1 Cơ sở dữ liệu NoSQL

NoSQL là một cơ sở dữ liệu phi quan hệ, cung cấp một cơ chế để lưu trữ và truy xuất dữ liệu được mô hình hóa khác với các quan hệ bảng được sử dụng trong các cơ sở dữ liệu truyền thống là các cơ sở dữ liệu kiểu quan hệ.

2.6.2 MongoDB

MongoDB là một chương trình cơ sở dữ liệu NoSQL. MongoDB sử dụng mô hình Database - Collection - Document thay thế mô hình cơ sở dữ liệu dùng Bảng truyền thống bằng các Document có cấu trúc linh hoạt.

Các khái niệm cơ bản trong MongoDB:

Document

Đơn vị lưu trữ thông tin trong MongoDB, tương ứng với Hàng (Row) trong hệ cơ sở dữ liệu quan hệ.

Document có cấu trúc gồm các cặp Khóa - Giá trị (key-value), tương tự cấu trúc JSON. Document trong cùng một cấp không nhất thiết có cấu

trúc giống nhau.

Collection

Collection là tập hợp một nhóm các Document, tương đương với Bảng (Table). Các Document trong một Collection có thể có cấu trúc khác nhau, tuy nhiên thông thường tất cả các Document đó biểu diễn các dữ liệu giống nhau liên quan tới nhau.

Database

Database là lớp chứa vật lý cho các Collection. Mỗi Database được thiết lập riêng một hệ thống quản lý tập tin. Mỗi máy chủ MongoDB thông thường cho phép nhiều Database.

2.7 Giới thiệu về GitHub và GitHub Actions

2.7.1 GitHub

GitHub là một dịch vụ cung cấp kho lưu trữ mã nguồn Git dựa trên nền web cho các dự án phát triển phần mềm. GitHub cung cấp kho lưu trữ miễn phí nếu các dự án đó làm mã nguồn mở. Tính đến tháng 4 năm 2016, GitHub có hơn 14 triệu người sử dụng với hơn 35 triệu kho mã nguồn, là máy chủ chứa mã nguồn lớn nhất trên thế giới.

2.7.2 GitHub Actions

Tổng quan

GitHub Actions là một dịch vụ mới của GitHub, GitHub Actions giúp tự động hóa các tác vụ lặp đi lặp lại trong vòng đời phát triển phần mềm.

GitHub Actions hoạt động theo hướng sự kiện (event-drivent), nghĩa là ta có thể thực thi một chuỗi câu lệnh sau khi một sự kiện cụ thể đã xảy ra. Ví dụ, mỗi khi ai đó tạo một pull request trên một repository, ta có thể chạy một cách tự động một câu lệnh để thực thi một tập lệnh để kiểm thử phần mềm.

Các thành phần chính

- Workflow: Một thủ tục chạy tự động được thêm vào repository. Workflow được cấu thành từ một hoặc nhiều job và có thể lên lịch hoặc kích hoạt bởi một event.
- Event: Một sự kiện là một hoạt động có thể kích hoạt một workflow.
- Job: Một tập hợp các step. Mặc định, một workflow bao gồm nhiều job sẽ thực thi chúng song song nhau. Tuy nhiên chúng vẫn có thể được cấu hình để chạy tuần tự.
- Step: Tương tự job nhưng nhỏ hơn. Một step là một tác vụ để thực thi các câu lệnh trong một job.
- Runner: Runner là một máy chủ ảo, điều khiển bởi GitHub, lắng nghe các job và thực thi chúng. Một runner được thực thi trên môi trường máy chủ ảo trên nền hệ điều hành Ubuntu Linux, Microsoft Windows, và macOS (có thể chỉ định và thay đổi).

Một trong những tính năng chính của GitHub Actions, đó là một workflow có thể được kích hoạt bằng cách lên lịch. Ví dụ, ta có thể cấu hình một workflow chạy định kỳ mỗi giờ, mỗi 3 giờ hay hàng ngày. Điều này khiến GitHub Actions đáp ứng rất tốt một trong các yêu cầu phổ biến của các ứng dụng máy chủ hiện nay: thực thi tác vụ được lên lịch. Ví dụ, một ứng dụng có thể cần gọi một API mỗi 5 phút, hay gửi email báo cáo vào mỗi buổi tối. Lợi ích khi sử dụng GitHub Actions để triển khai mô đun thực thi code lặp lại theo chu kỳ (cron):

- Hoàn toàn miễn phí.
- Không giới hạn số tác vụ.
- Thời gian giới hạn mỗi chu kỳ chạy là 6 giờ.
- Bảo mật thông tin mật với biến môi trường.

2.8 Hệ thống tổng hợp thông tin trên báo chí

Hệ thống tổng hợp thông tin trên báo chí, hay Hệ thống tổng hợp tin tức, là một dịch vụ web được tạo ra nhằm mục đích thu thập các nguồn tin tức, sau đó phân tích, chọn lọc và trình bày chúng theo thói quen và nhu cầu nắm bắt thông tin của người dùng. Do số lượng các trang tin tức quá lớn, người dùng có nhu cầu tổng hợp những tin tức nóng hổi nhất từ tất cả các trang báo đó và tiếp thu một cách liền mạch. Đó chính là lý do ra đời các hệ thống tổng hợp tin tức.

2.8.1 Hệ thống tổng hợp tin tức dựa trên công nghệ RSS Feeds

Định dạng tập tin RSS

RSS (Resource Description Framework) Site Summary là một tập tin dựa trên XML dùng trong việc chia sẻ tin tức Web được dùng bởi những nhà cung cấp tin tức và blog. Những nhà cung cấp tin tức này cho phép các trang web khác tổng hợp những tiêu đề tin tức "được chia sẻ" hay cung cấp các tóm tắt ngắn gọn của các bản tin chính dưới nhiều hình thức khác nhau.

Một tập tin RSS thường chứa những thông tin sau:

- Tên và mô tả của kênh RSS.



Hình 2.2: Biểu tượng nhận diện một trang web có hỗ trợ RSS

- Tiêu đề của mỗi bài viết.
- Tóm tắt nội dung bài viết.
- Đường dẫn URL dẫn tới bài viết gốc.
- Đường dẫn của hình thu nhỏ bài viết.

Sau đây là một tập tin RSS mẫu:

```
1 <rss version="2.0">
2 <channel>
3 <title>Sample Feed</title>
4 <description>RSS is a fascinating technology. The uses for RSS are expanding daily
. </description>
5 <link>http://www.tindiaphuong.org/hot-news-1.htm</link>
6 <category>Entertainment</category>
7 <language>en-us</language>
8 <lastBuildDate>Tue, 05 Jan 2021 13:39:14 -0400</lastBuildDate>
9 <pubDate>Tue, 19 Oct 2004 13:38:55 -0400</pubDate>
10
11 <generator>TDP Generator</generator>
12 <image>
13 <url>https://www.tindiaphuong.org/hot-news-2.htm</url>
14 <title>Sample Feed</title>
15 <link>http://www.feedforall.com/industry-solutions.htm</link>
16 <description>Sample Feed</description>
17 <width>48</width>
18 <height>48</height>
19 </image>
20 <item>
21 <title>RSS Restaurants</title>
```

```

22 <description>FeedForAll helps Restaurant's communicate with customers. Let your
   customers know the latest specials or events.</description>
23 <link>http://www.feedforall.com/restaurant.htm</link>
24 <category domain="www.dmoz.com">Computers/Software/Internet/Site Management/
   Content Management</category>
25 <comments>http://www.feedforall.com/forum</comments>
26 <pubDate>Tue, 19 Oct 2004 11:09:11 -0400</pubDate>
27 </item>
28 </channel>
29 </rss>

```

Listing 2.1: Cấu trúc một tập tin RSS

Trên các trang web hỗ trợ RSS, các RSS feeds thường được liên kết bằng một hình chữ nhật màu cam, có thẻ kèm theo các ký tự XML hay RSS.

Một chương trình thu thập RSS có thể kiểm tra một trang web có hỗ trợ RSS hay không và, nếu có, hiển thị những bài viết cập nhật nhất mà nó tìm thấy từ trang web đó.

Mặc dù một hệ thống tổng hợp tin tức sử dụng công nghệ RSS có thể được xây dựng một cách nhanh chóng, tuy nhiên chúng có một số nhược điểm khiến cho công nghệ này không còn được ưa chuộng ngày nay:

- Hệ thống chỉ có thể thu thập tin tức từ các trang có hỗ trợ công nghệ RSS, dẫn tới số lượng nguồn tin thu thập được hạn chế.
- Không lấy được nội dung bài viết.

2.8.2 Hệ thống tổng hợp tin tức bằng cách cào dữ liệu các trang tin tức

Khái niệm cào dữ liệu

Cào dữ liệu (tiếng Anh: Data crawling) là một kỹ thuật thu thập dữ liệu trên Internet phổ biến hiện nay. Một trình thu thập dữ liệu web ("Web crawler") được xây dựng để thực hiện chức năng cào dữ liệu. Được ví như

một con bọ, công cụ cào dữ liệu bắt đầu bằng việc ghé thăm một danh sách trang web có đường dẫn (URL) được chuẩn bị trước. Các trang web này được ví như những "hạt giống" để "con bọ" định kỳ ghé thăm và thu thập.

Trình thu thập dữ liệu web

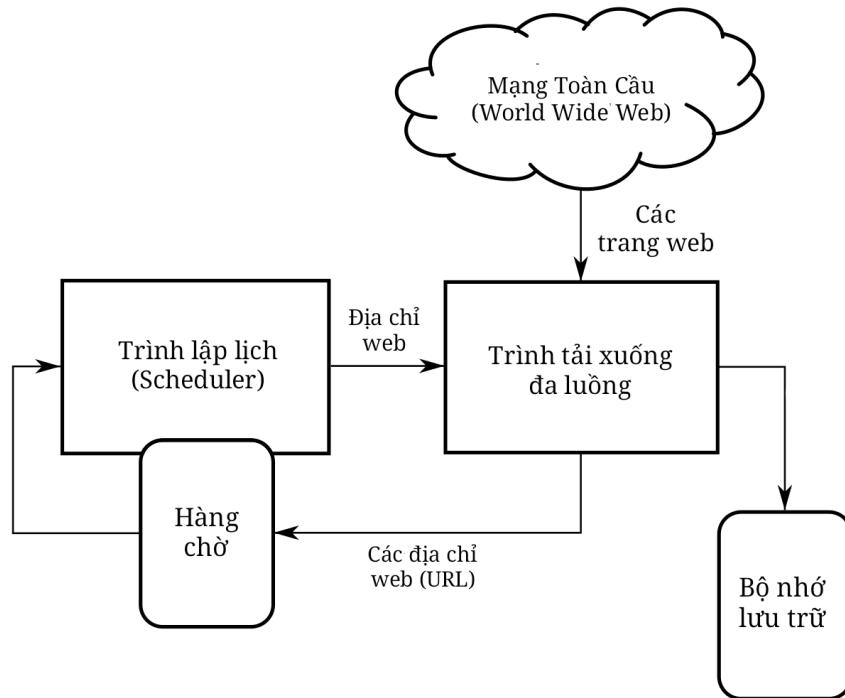
Trình thu thập dữ liệu web là một chương trình phần mềm có khả năng tự động lấy dữ liệu gồm thông tin, ảnh, video,... trên mạng Internet. Thông thường, trình thu thập dữ liệu thường chạy trên một máy chủ (server), chịu trách nhiệm lấy thông tin, trích xuất những thông tin mà người sử dụng cần, đồng thời cũng tìm những đường dẫn mới có từ các thông tin có được và tự động truy cập vào những đường dẫn mới đó.

Mô hình một trình thu thập dữ liệu web đơn giản gồm:

- 1. Chọn một, hoặc các đường dẫn khởi đầu.
- 2. Sử dụng giao thức HTML để tải trang web.
- 3. Trích xuất ra nội dung cần thiết và các siêu liên kết mới và thêm chúng vào hàng chờ.
- 4. Lặp đi lặp lại các bước 2, 3.

Các vấn đề sẽ gặp phải khi xây dựng một trình thu thập dữ liệu web:

- Giới hạn thời gian: Nếu máy chủ của một trang không trả lời thì chương trình sẽ bị đóng băng. Vì vậy cần xử lý trường hợp máy chủ không trả lời sau một khoảng thời gian nhất định.
- Lên kế hoạch truy cập một cách hợp lý: Do chương trình sẽ liên tục truy xuất các trang web để cập nhật dữ liệu mới nhất, vì thế cần phải quản lý tần suất truy cập một trang, ví dụ chỉ truy xuất lại sau 3 đến 6 tiếng.



Hình 2.3: Mô hình một trình thu thập dữ liệu web đơn giản

- Không truy cập lại trang web đã xử lý xong: Nếu không xử lý vấn đề này, chương trình sẽ bị rơi vào vòng lặp vĩnh viễn. Vì thế cần phải xây dựng phương pháp đánh dấu những trang đã xử lý. Đơn giản nhất là lưu lại địa chỉ web (URL) của chúng. Trước khi thêm vào hàng chờ một địa chỉ web mới thì so sánh với những địa chỉ đã xử lý trước đó.
- Chiến thuật lưu trữ hợp lý: Thông thường sau khi thu thập sẽ cần lưu trữ một lượng dữ liệu khổng lồ và ngày qua ngày sẽ càng phình to ra. Vì vậy yêu cầu một bộ nhớ lưu trữ có tốc độ đọc ghi cao, hiệu suất lớn và dễ dàng mở rộng.

2.8.3 Sử dụng trình thu thập dữ liệu web để thu thập tin tức

Các hệ thống tổng hợp tin tức về bản chất là sử dụng hệ thống thu thập dữ liệu web cho mục đích thu thập tin tức.

Chương 3

Thiết kế

3.1 Giải pháp tổng quát

Giải pháp tổng quát cho hệ thống tổng hợp tin tức địa phương được chia làm 3 tác vụ chính:

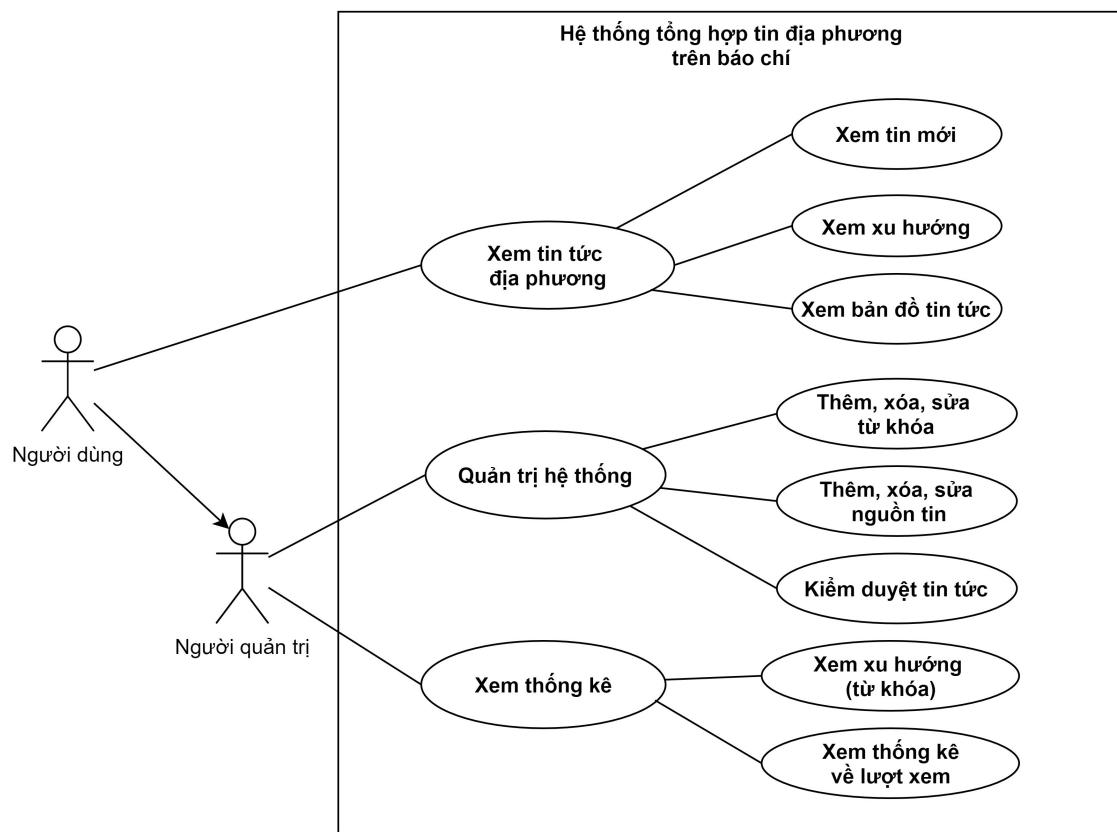
- Tổng hợp tin tức: Xây dựng hệ thống tự động lấy tin tức từ các nguồn tin và lưu vào cơ sở dữ liệu.
- Lọc và phân tích tin tức địa phương: Lọc ra các tin tức địa phương từ các tin tức tổng hợp được. Sau đó chạy các tác vụ phân tích các tin tức địa phương.
- Xây dựng API cung cấp dữ liệu để hiển thị tin tức lên trình duyệt web.

3.2 Thiết kế hệ thống

Hệ thống hoạt động dựa trên kiến trúc Client-Server.

Máy khách trong hệ thống "Ứng dụng tổng hợp thông tin địa phương trên báo chí" là trình duyệt web của người dùng.

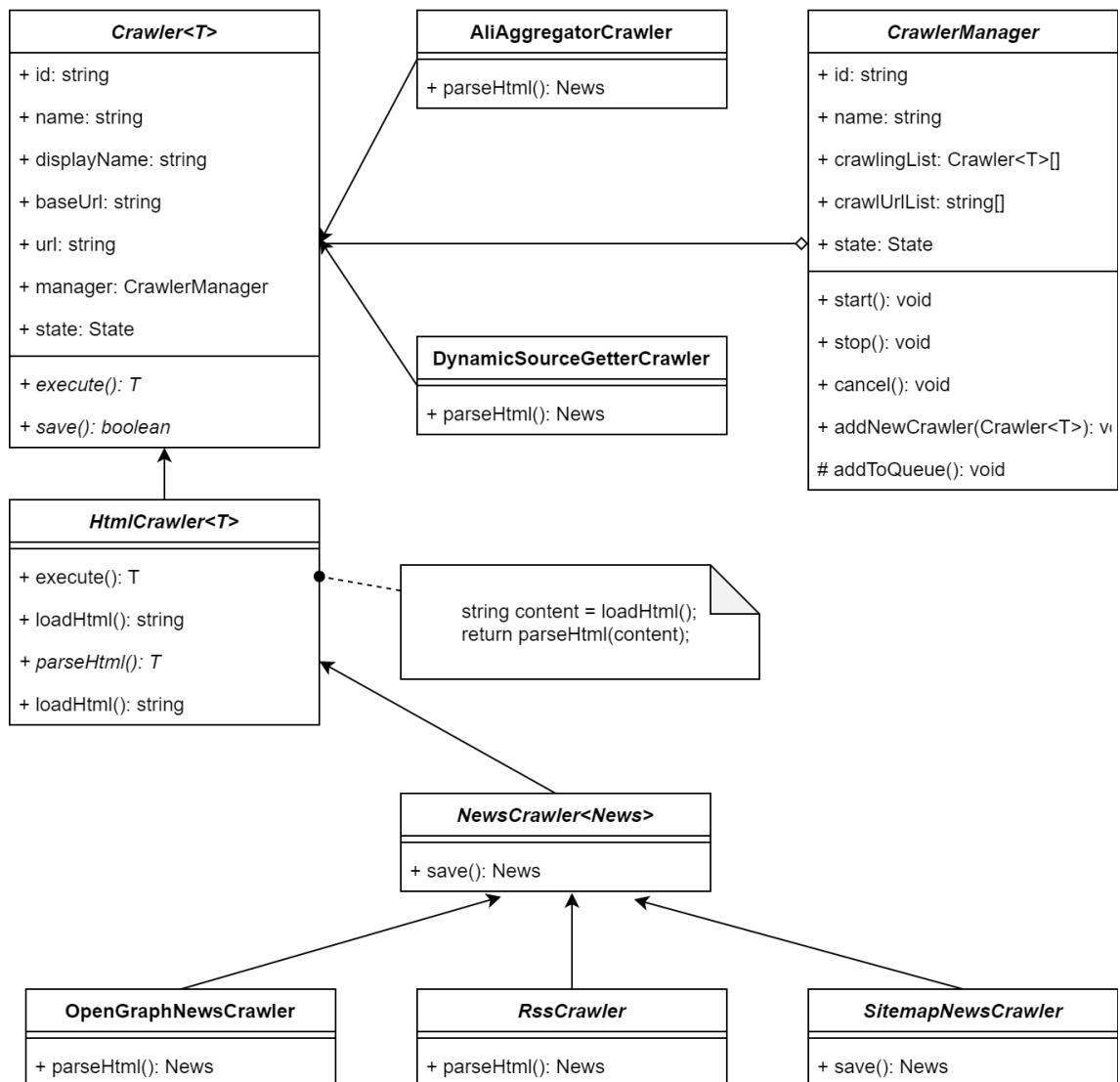
Sơ đồ tình huống sử dụng (Use Cases Diagram)



Hình 3.1: Sơ đồ tình huống sử dụng

Máy chủ sử dụng hệ thống các dịch vụ máy chủ của các nhà cung cấp. Các dịch vụ máy chủ đảm bảo duy trì sự hoạt động cho mô-đun "Tổng hợp và phân tích tin tức" và mô-đun "Dịch vụ RESTful API".

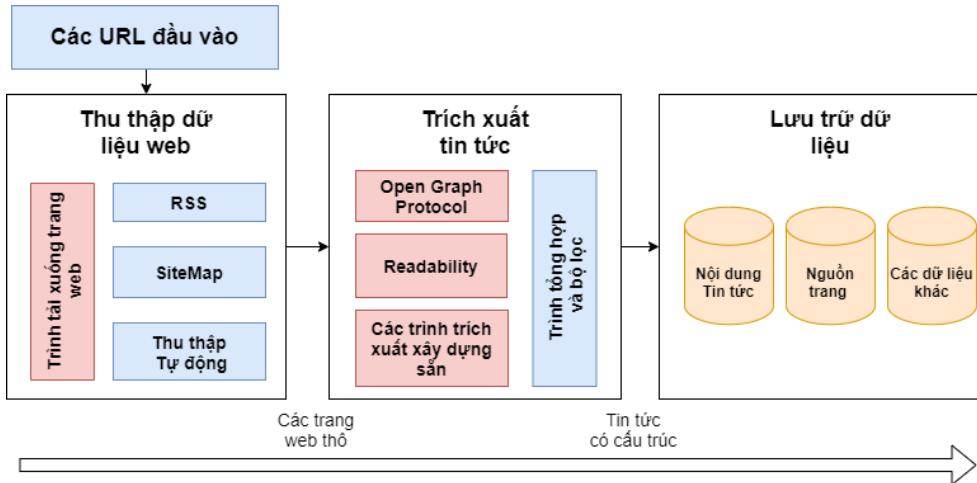
Sơ đồ lớp (Class Diagram)



Hình 3.2: Sơ đồ tình huống sử dụng

3.3 Tổng hợp tin tức

Tổng hợp tin tức từ các nguồn tin báo chí là chức năng chính yếu của hệ thống. Chúng tôi sử dụng kết hợp 2 giải pháp về tổng hợp tin tức đã nêu, kết hợp tổng hợp tin tức bằng RSS, Sitemap và trình thu thập dữ liệu web tự động.



Hình 3.3: Thu thập và trích xuất tin tức

3.3.1 Thu thập dữ liệu web từ các URL đầu vào

Tiến hành thu thập các địa chỉ web của các nguồn tin tức phổ biến hiện nay. Các địa chỉ của các nguồn tin tức này thường cố định và được các trang tin cập nhật theo chu kỳ, nội dung của chúng thường chứa các đường dẫn dẫn tới các bài viết chi tiết.

Mục tiêu của bước "Thu thập dữ liệu web" là lấy toàn bộ đường dẫn tới bài viết chi tiết, từ nội dung trang web nguồn, sau đó lần lượt tải về nội dung của chúng để trình trích xuất tin tức xử lý.

Các trang web nguồn được phân loại thành một trong ba loại chính và xử lý riêng biệt:

- Nếu nội dung thuộc định dạng RSS: Sử dụng trình đọc RSS phân tích để thu được danh sách tin tức và đường dẫn tới các tin tức đó.

- Nếu nội dung thuộc định dạng Sitemap: Sử dụng trình đọc Sitemap phân tích để thu được danh sách tin tức và đường dẫn tới các tin tức đó.
- Nếu nội dung không thuộc 2 định dạng nêu trên, Sử dụng trình thu thập tự động.

3.3.2 Trình trích xuất tin tức

Trình trích xuất tin tức là trái tim của một hệ thống tổng hợp tin tức. Một trình trích xuất tin tức cần đảm bảo hoạt động ổn định, ít sai sót nhất có thể. Vì vậy, chúng tôi xây dựng trình trích xuất kết hợp 3 giải pháp có sẵn để đảm bảo tin tức được trích xuất ra một cách chính xác:

3.3.3 Trích xuất dựa trên giao thức Open Graph

Giao thức Open Graph (Open Graph protocol) [8] là một giao thức dùng để giao tiếp giữa trang web và các mạng xã hội. Được tạo ra bởi mạng xã hội Facebook, ban đầu chuẩn này được tạo ra để giúp Facebook hiểu và tóm tắt được nội dung trang web chỉ qua đường dẫn được chia sẻ. Hiện nay, Open Graph Protocol là một tiêu chuẩn phổ biến và hầu như tất cả các trang web hiện nay đều hỗ trợ.

Về cơ bản, chuẩn này yêu cầu quản trị viên trang web "chèn" vào trang web các thẻ meta theo cấu trúc được quy định.

```

1 <html prefix="og: https://ogp.me/ns#">
2 <head>
3 <title>The Rock (1996)</title>
4 <meta property="og:title" content="The Rock" />
5 <meta property="og:type" content="video.movie" />
6 <meta property="og:url" content="https://www.imdb.com/title/tt0117500/" />
7 <meta property="og:image" content="https://ia.media-imdb.com/images/rock.jpg" />
8 ...
9 </head>
10 ...

```

Listing 3.1: Mã nguồn một trang web hỗ trợ Open Graph Protocol

Các trang tin tức dưới dạng bài viết hoàn toàn có thể phân tích được nội dung nếu nó hỗ trợ chuẩn OG Protocol. Vì vậy, chúng tôi đã sử dụng một giải pháp là phân tích nội dung trang sử dụng trình đọc OG Protocol để trích xuất nội dung tin tức. Khi trích xuất tin tức bằng cách này, có thể thu được các thông tin khá đầy đủ gồm:

- Tiêu đề bài viết
- Tóm tắt nội dung
- đường dẫn ảnh thu nhỏ
- Ngày xuất bản
- Tác giả bài viết

Trình trích xuất dựa trên giao thức Open Graph có ưu điểm là độ chính xác cao do được quản trị trang web khai báo thủ công. Tuy nhiên, trình trích xuất không thể kiểm tra mức độ chính xác được khai báo. Ngoài ra, cách này còn một nhược điểm là không thể trích xuất được phần nội dung của bài viết.

3.3.4 Trích xuất sử dụng thư viện Readability

Readability [9] (Trình đọc trang) là thư viện mã nguồn mở có khả năng loại bỏ các thẻ HTML, chỉ giữ lại nội dung bài viết. Readability thường được sử dụng trong các trình duyệt web dưới chức năng "Chế độ đọc" hay "Chế độ đọc giả" (Reader mode).

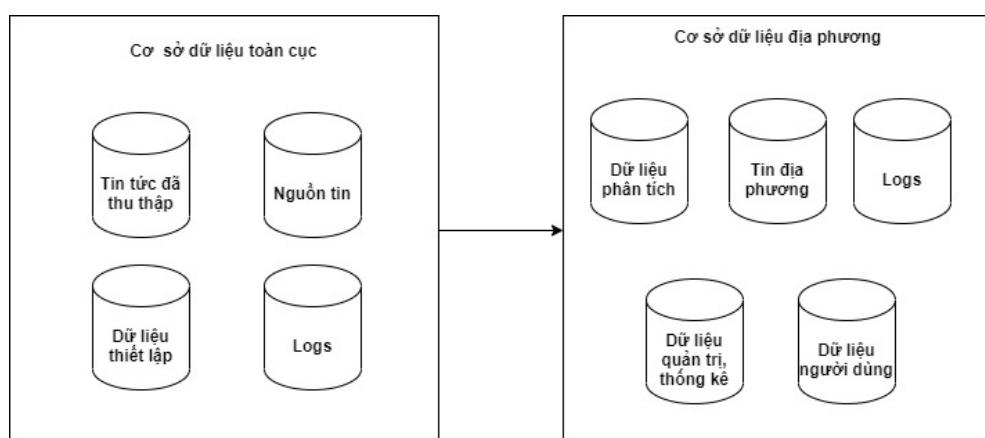
Trình trích xuất sử dụng Readability ngoài lấy được thông tin gồm tiêu đề, tóm tắt nội dung... còn có thể trích xuất được nội dung của bài viết. Tuy nhiên, tin tức trích xuất từ giải pháp này có mức độ chính xác tương đối và không chính xác bằng trình trích xuất Open Graph protocol.

3.3.5 Xây dựng trình trích xuất sẵn cho các nguồn tin tức phổ biến

Đối với các nguồn tin phổ biến thì chúng tôi xây dựng sẵn các trình trích xuất dành riêng cho các trang này, giải pháp này có ưu điểm là chính xác và lấy được toàn bộ thông tin cũng như nội dung của bài viết, nhưng nhược điểm là chỉ hỗ trợ các nguồn tin phổ biến và ổn định (ít thay đổi định dạng trang web).

Trình tổng hợp tin tức sau đó sẽ xem xét kết quả từ cả 3 bước kể trên, kết hợp lại để cho ra kết quả phù hợp nhất. Ngoài ra, trình tổng hợp tin tức còn thực hiện lọc để loại bỏ các tin tức trùng lặp, không hợp lệ.

3.3.6 Lưu trữ dữ liệu



Hình 3.4: Lưu trữ dữ liệu

Dữ liệu được lưu trữ dưới định dạng có cấu trúc để hệ thống có thể truy cập sử dụng dễ dàng. Chúng bao gồm dữ liệu tin tức đã thu thập, dữ liệu tin tức địa phương, dữ liệu thiết lập và dữ liệu người dùng.

Cơ sở dữ liệu được chia thành hai phần chính:

- Cơ sở dữ liệu toàn cục: Lưu trữ dữ liệu tin tức đã thu thập được. Các dữ liệu tin tức này dùng chung cho toàn bộ các địa phương.

- Cơ sở dữ liệu địa phương: Mỗi địa phương có các cơ sở dữ liệu riêng, bao gồm dữ liệu tin tức địa phương, dữ liệu sau phân tích và các thiết đặt người dùng.

Cơ sở dữ liệu được sử dụng trong ứng dụng là cơ sở dữ liệu MongoDB.

3.4 Lọc tin tức địa phương

Các bài viết được xem là tin địa phương của một địa phương, khi:

- Tiêu đề, tóm tắt hoặc nội dung bài viết chứa tên hoặc vị trí địa lý của địa phương. Ví dụ: Bài viết có chứa từ khóa "Quận 9" được xác định là tin địa phương của Quận 9.
- Tiêu đề, tóm tắt hoặc nội dung bài viết chứa tên hoặc vị trí địa lý của địa phương nằm trong địa phương đó. Ví dụ: Bài viết có chứa từ khóa "phường Phước Long A" được cho là tin địa phương của Quận 9.
- Tiêu đề, tóm tắt hoặc nội dung bài viết chứa tên riêng của các địa danh, công trình nổi tiếng.. của địa phương đó.

3.5 Tin chính

Tin chính gom nhóm các tin tức cùng chủ đề, cho phép người dùng tập trung vào một chủ đề và dễ dàng tìm đọc các tin liên quan. Đây chính là bài toán *Dánh giá sự tương đồng* và *Gom nhóm văn bản* mà chúng tôi đã đề cập tại chương 2.

Các bước lọc tin chính:

- *Lọc tin chính* xảy ra ngay sau quá trình *Lọc tin tức địa phương*.
- Dánh giá sự tương đồng của hai tin bất kỳ.

- Gom nhóm các tin có sự tương đồng bằng các thuật toán DBSCAN.
- Sắp xếp thứ tự các nhóm dựa trên mức độ phổ biến.
- Lưu vào cơ sở dữ liệu.

3.6 Xu hướng tin tức

Hệ thống thống kê sự xuất hiện (số lần xuất hiện) của các từ khóa phổ biến. Từ đó cho ra *Xếp hạng từ khóa*. *Xếp hạng từ khóa* là một bảng xếp hạng các từ khóa "nóng" nhất tại một thời điểm.

Các từ khóa này được thể hiện trong một giao diện đồ họa trực quan, cho phép người dùng nắm bắt được những gì đang diễn ra tại địa phương của mình.

Các tin tức "nóng" là các tin tức có tiêu đề, tóm tắt hoặc nội dung chứa các từ khóa phổ biến. Do đó, từ dữ liệu *Xếp hạng từ khóa*, hệ thống có thể xếp thứ tự các tin theo độ "nóng" của chúng một cách chính xác nhất.

3.7 Bản đồ tin tức

Tin tức sau khi tổng hợp được thống kê lại theo các khu vực. Sau đó, chúng được hiển thị lại thành mật độ tin tức của một khu vực cụ thể. Các thông tin mà bản đồ tin tức đem lại là:

- Các địa phương thuộc khu vực đang xét. Ví dụ như quận có các phường là Phường Hiệp Phú, Phường Long Bình, Phường Long Phước,...
- Một độ tin tức được thể hiện dưới dạng vòng tròn độ rộng theo mật độ tin thuộc địa phương đó.

- Bản đồ cung cấp khả năng tùy chỉnh theo đường, hoặc theo các thông số khác.
- Bản đồ sử dụng API cấp bởi Google Map có độ chính xác cao.
- Có thể xem các tin tức xảy ra tại địa điểm khi chọn vào các địa điểm đó trên bản đồ.

3.8 Quản trị thống kê

Nhà quản trị địa phương có thể thêm, loại các từ khóa để cải thiện độ chính xác của thuật toán. Nhà quản trị địa phương cũng có thể ngăn chặn một nguồn tin, kiểm duyệt hoặc loại bỏ các tin tức không phù hợp.

Ngoài ra, nhà quản trị địa phương còn có thể xem thống kê chi tiết về lượt xem, các chủ đề được xem nhiều, xem thống kê về xu hướng tin tức (chi tiết hơn so với xu hướng phiên bản thông thường).

Chương 4

Cài đặt

4.1 Môi trường thực nghiệm

Ứng dụng Back-end được triển khai trên máy chủ, có trách nhiệm cung cấp dữ liệu cho ứng dụng giao diện web (Front-end). Ngoài ra ứng dụng Back-end còn đảm bảo thực thi theo chu kỳ mô-đun tổng hợp và phân tích tin tức.

Ứng dụng Back-end xây dựng trên công nghệ Node.js, ngôn ngữ TypeScript trên trình biên tập mã Microsoft Visual Studio Code 1.53.2

Các thư viện được sử dụng:

Tên thư viện	Mô tả, chức năng
axios	Tải về nội dung web
@mozilla/readability	Trích xuất nội dung tin tức khỏi trang web
article-parser	Trích xuất nội dung tin tức khỏi trang web
rss-parser	Phân tích nội dung tập tin RSS
sitemapper	Phân tích nội dung tập tin sitemap
vntk	Thư viện xử lý ngôn ngữ tự nhiên (Tiếng Việt)
natural	Thư viện xử lý ngôn ngữ tự nhiên

leven	Tính toán sự tương đồng hai chuỗi
string-similarity	Tính toán sự tương đồng hai chuỗi
density-clustering	Gom nhóm các nội dung tương tự nhau
set-clustering	Gom nhóm các nội dung tương tự nhau
bcryptjs	Mã hóa mật khẩu
dotenv	Khai báo biến môi trường
mongodb	Driver cho phép ứng dụng Node.js truy cập các API của MongoDB
mongoose	Khai báo và định kiểu dữ liệu cho MongoDB
passport	Thư viện hỗ trợ xác thực người dùng

Bảng 4.1: Các thư viện được sử dụng

Các thư viện ở trên có thể được cài đặt vào bất cứ dự án Node.js nào thông qua lệnh:

```

1 npm install $TEN_THU_VIEN
2 npm run build

```

Listing 4.1: Cài đặt một thư viện vào hệ thống sử dụng Node.js

4.2 Mô đun tổng hợp và phân tích tin tức

Mô-đun *tổng hợp và phân tích tin tức* được triển khai dưới dạng tác vụ bởi dịch vụ GitHub Actions. Tác vụ được lặp lại theo chu kỳ mỗi giờ và có thể theo dõi tại địa chỉ https://github.com/ali-kit/ali_be_ts/actions

```

1 name: News Crawling and Analytics
2
3 on:
4   workflow_dispatch:
5     schedule:
6       - cron: "0 * * * *"
7

```

```

8 jobs:
9   crawl-news:
10     name: Crawl news
11     runs-on: ubuntu-latest
12
13   steps:
14     - uses: actions/checkout@v2
15     - name: Use Node.js
16       uses: actions/setup-node@v1
17       with:
18         node-version: '12.x'
19     - uses: actions/cache@v2
20       with:
21         path: ~/.npm
22         key: ${{ runner.os }}-node-${{ hashFiles('**/package-lock.json') }}
23         restore-keys: |
24           ${{ runner.os }}-node-
25     - name: Install Dependencies & Build
26       run: |
27         npm install
28         npm run build
29     - name: Crawl News
30       run: |
31         npx ts-node -r tsconfig-paths/register ./src/scripts/analyzer/index --env=production

```

Listing 4.2: Tập lệnh workflows để triển khai

4.3 Dịch vụ RESTful API

Bên cạnh mô-đun *tổng hợp và phân tích tin tức*, hệ thống yêu cầu một dịch vụ RESTful API nhằm cung cấp nguồn dữ liệu cho ứng dụng giao diện Front-end.

Dịch vụ cung cấp RESTful API được triển khai trên máy chủ Heroku.

All workflows			
Showing runs from all workflows			
<input type="text"/> Filter workflows			
784 workflow runs	Event ▾	Status ▾	Branch ▾
Actor ▾			
⌚ News Crawling and Analytics News Crawling and Analytics #559: Scheduled	⌚ 33 minutes ago	In progress	...
⌚ News Crawling and Analytics News Crawling and Analytics #558: Scheduled	⌚ 1 hour ago	56m 58s	...
⌚ News Crawling and Analytics News Crawling and Analytics #557: Scheduled	⌚ 2 hours ago	1h 2m 14s	...
⌚ News Crawling and Analytics News Crawling and Analytics #556: Scheduled	⌚ 4 hours ago	1h 5m 52s	...
⌚ News Crawling and Analytics News Crawling and Analytics #555: Scheduled	⌚ 5 hours ago	1h 0m 50s	...

Hình 4.1: GitHub Actions thực thi tác vụ theo chu kỳ

4.4 Giao diện Front-end

Front-end chịu trách nhiệm hiển thị nội dung (sử dụng dữ liệu được cấp bởi dịch vụ RESTful API). Ứng dụng được xây dựng trên nền tảng Node.js và thư viện giao diện React.js, ngôn ngữ phát Javascript. Ứng dụng được phát triển trên trình biên tập mã Microsoft Visual Studio Code 1.53.2

Ứng dụng Front-end được triển khai bằng dịch vụ GitHub Pages của GitHub tại địa chỉ <https://tindiaphuong.github.io>

4.4.1 Tin chính

Trang Tin chính, là trang chủ của ứng dụng. Trang tin chính hiển thị danh sách tin tức được sắp xếp theo thời gian và mức độ phổ biến. Nếu các tin tức liên quan với nhau (cùng nói về một chủ đề và có mức độ tương đồng cao), chúng sẽ được gom thành một nhóm. Người dùng có thể xem nhấn chọn *Xem toàn cảnh* để xem tất cả các tin trong một nhóm.

Tin địa phương
Quận 9

Tin chính Xu hướng Tin mới Bản đồ tin tức Quản trị (dành cho quản trị viên)

Tin Chính

Zing • 3 ngày trước
Chi số UV tại TP.HCM tiếp tục ở ngưỡng gây hại cao
Hôm nay và 2 ngày trước, chỉ số UV tại TP.HCM duy trì ngưỡng có nguy cơ gây hại rất cao - 10 đơn vị. Trong khi đó, trời se lạnh vào đêm và sáng sớm.

- Tia cực tím đạt cực đại 3 ngày liên tiếp tại TP.HCM
- Chi số UV tại TP.HCM ở mức gây hại rất cao
- Hôm nay chỉ số tia UV tại TP.HCM đạt cực đại

Dân Sinh • 3 ngày trước
Bộ Xây dựng: Giá nhà ở riêng lẻ, đất nền nhiều khu vực tăng 20-50%
Bộ Xây dựng cho biết, năm 2020, giá nhà ở riêng lẻ, đất nền vẫn có xu hướng tăng hơn so với năm 2019. Tuy nhiên biến độ tăng giá rất khác nhau giữa các địa phương cũng như tại từng khu vực cụ thể.

VietnamPlus • 2 ngày trước
Đồng Nai: Tạm giữ hình sự chủ quán karaoke tổ chức sử dụng ma túy
Ranh sảnh 21/2 Cảnh sát điều tra Công an Rạch Giá thành phố Rào Hòa bắt người kiểm tra quán karaoke Non Linh nhất hiện 17 đối tượng sử dụng ma túy

Hình 4.2: Giao diện trang chủ Tin chính

4.4.2 Xu hướng

Xu hướng mang tới cho người dùng cái nhìn tổng quan về những gì đang xảy ra trên địa phương. Giao diện trang Xu hướng gồm 2 phần:

- Phân mục trái: hiển thị danh sách tin tức phổ biến hiện nay.
- Phân mục phải: hiển thị *đám mây từ khóa*. Với đám mây từ khóa, các chủ đề (từ khóa) càng "nóng" thì chúng được hiển thị càng lớn. Người dùng khi nhấn chọn một từ khóa, một trang sẽ mở ra và họ có thể xem toàn bộ tin tức liên quan tới nó.

4.4.3 Bản đồ tin tức

Bản đồ tin tức thể hiện trực quan phân bố tin tức bên trong địa phương (phường, xã). Các tin tức của cùng một khu vực nhỏ được gom nhóm lại và được thể hiện bằng một chấm tròn trên bản đồ. Càng nhiều tin tức trong khu vực đó thì kích thước chấm tròn càng lớn và ngược lại.

Tin địa phương
Quận 9

[Đăng nhập](#)

[Tin chính](#) [Xu hướng](#) [Tin mới](#) [Bản đồ tin tức](#) [Quản trị \(dành cho quản trị viên\)](#)

Xu Hướng

 Nhân Dân • 19 giờ trước

Nổ lốp ô-tô khiến chủ tiệm thiệt mạng

Chiều 24.2, trên địa bàn huyện Phú Riềng xảy ra một vụ nổ lốp xe ô-tô khiến ông Khuất Đức Hiển (41 tuổi), ngụ phường Long Phước, thị xã Phước Long, tỉnh Bình Phước, chủ tiệm sửa chữa, hàn tiên ô-tô Chí Thiện ở thôn Tân Lực, xã Bù Nhô, huyện Phú Riềng tử vong.

 Giao Thông • 22/02/2021 17:16

Những công trình nghìn tỷ nào sẽ là cú hích, 'thắp sáng' TP Thủ Đức?

Ha tầng giao thông là một ưu điểm nổi bật của khu Đông TP.HCM với hàng loạt dự án hạ tầng trọng điểm đã và đang được triển khai.

 Zing • 3 ngày trước

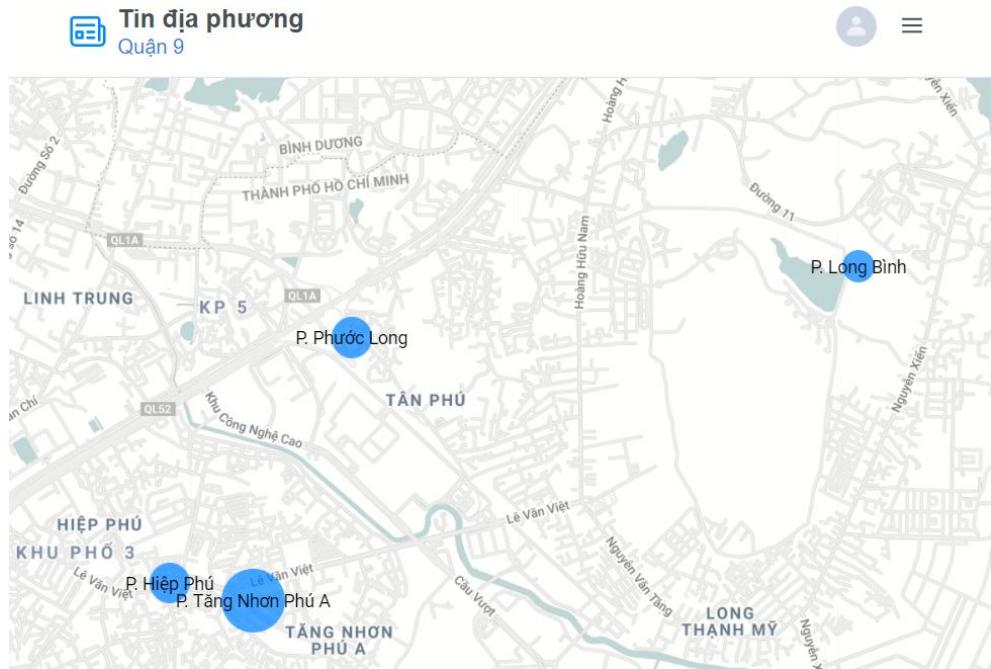
Chi số UV tại TP.HCM tiếp tục ở ngưỡng gây hại cao

Hôm nay và 2 ngày tới, chỉ số UV tại TP.HCM duy trì ngưỡng có nguy cơ gây hại rất cao - 10 đơn vị. Trong khi đó, trời se lạnh vào đêm và sáng sớm.

Tăng Nhơn Phú A ↓ Ca bệnh ↓ Ma túy ↑
Quận 12 ↑ Nhau ↑ Gò Vấp ↑ Covid-19 ↓
Dich ↑ Long Trường ↑ Truy ↓ Trần Văn Danh ↑
Cách ly ↑ TP.HCM ▲
Công an TP Thủ Đức ↑ Tết ▲ Tân Bình ↑
Long Bình Tân ↓ Người dân ồng ↓ Nhà cấp bón ↓
Xét nghiệm ↓ Quán karaoke ↓ Lắc ↓
Nguyễn Thành Hưng ↓
COVID-19 ▼

Thủ Đức ▲ Tử vong ↓
VIP ↓ Vườn nướng Mái đỏ ↓
Sân bay quốc tế Tân Sơn Nhất ↓ HCDC ▼
Trần Trung Nam ↑ Đồng Nai ↑ Nhiêm ↓
Đường Lê Văn Việt ↓ Lấy mẫu ↓ Tây Hòa ↑
Phước Long A ↑ Trung Mỹ Tây ↑
Công an TP Biên Hòa ↑ Trái phép ↑ Quán ăn ↑

Hình 4.3: Giao diện Xu hướng



Hình 4.4: Giao diện Bản đồ tin tức

Chương 5

Kết luận

5.1 Kết quả đạt được

Khóa luận *Ứng dụng tổng hợp thông tin địa phương trên báo chí* đã trình bày các kiến thức cơ bản về một hệ thống tổng hợp tin tức và phát hiện tin địa phương. Khóa luận cũng trình bày các các giải pháp cũng như cài đặt ứng dụng thực tế.

Với kết quả đạt được tương đối khả quan, chúng tôi nhận thấy ứng dụng thực tế đã phần nào đáp ứng được những yêu cầu cơ bản nhất của một hệ thống tổng hợp thông tin địa phương trên báo chí. Bên cạnh đó, ứng dụng cũng có sẵn các chức năng nổi bật, có tiềm năng so sánh với các sản phẩm trên thị trường.

5.2 Hạn chế

Tuy nhiên, chúng tôi cũng nhận thấy ứng dụng tổng hợp thông tin địa phương trên báo chí còn tồn tại một số hạn chế sau:

Giao diện

- Giao diện web còn đơn giản, chưa thực sự hấp dẫn người dùng.

- Chưa phân hóa tin tức theo sở thích của người dùng.

Mô-đun tổng hợp và phân tích tin tức

- Thời gian của một chu kỳ chạy còn kéo dài: ứng dụng cần 60 phút để hoàn thành một chu kỳ.
- Kết quả phân tích còn một số sai sót: một số tin tức được cho là tin địa phương dù không liên quan gì đến địa phương đó.

Thời gian cho một chu kỳ tổng hợp còn dài Hiện tại, mô-đun tổng hợp và phân tích tin tức cần trung bình 60 phút để hoàn thành một chu kỳ chạy.

5.3 Hướng phát triển trong tương lai

Để phát triển hoạt động ứng dụng tổng hợp thông tin địa phương trên báo chí hơn nữa, định hướng phát triển của khóa luận trước mắt là khắc phục những hạn chế kể trên, gồm nâng cấp giao diện, cải thiện thuật toán tổng hợp và phân tích tin tức.

Ngoài ra, chúng tôi mong muốn mở rộng các địa phương được hỗ trợ ra cả nước, không chỉ giới hạn tại các địa phương mẫu.

Thời gian thực hiện khóa luận đã giúp chúng tôi tích lũy được rất nhiều kiến thức và kinh nghiệm quý giá:

- Cách triển khai một ứng dụng web hoàn chỉnh.
- Cách hoạt động và triển khai một hệ thống tổng hợp tin tức.
- Các kiến thức về xử lý ngôn ngữ tự nhiên.
- Kỹ năng lập trình Node.js, React.js và TypeScript.
- Kỹ năng quản lý dự án và làm việc nhóm.
- Khả năng tìm kiếm và đọc hiểu tài liệu.

- Kỹ năng viết tài liệu hợp lý, trình bày đẹp mắt và hài hòa.

Tài liệu tham khảo

Tiếng Việt

- [1] Tuổi Trẻ Online. *Báo cáo Hội nghị báo chí toàn quốc 2020*. URL: <https://tuoitre.vn/bao-chi-giam-ca-so-luong-lan-doanh-thu-20201231100236058.htm> (visited on 12/31/2020).

Tiếng Anh

- [2] Alexa. *Top Sites in Vietnam*. URL: <https://www.alexa.com/topsites/countries/VN>.
- [3] Christopher D.Manning Prabhakar Raghavan, Hinrich Schütze. *Introduction to Information Retrieval*, pp. 58-60. Addison-Wesley.
- [4] Keil, Jan Martin. *Efficient Bounded Jaro-Winkler Similarity Based Search*. URL: <https://fusion.cs.uni-jena.de/fusion/wp-content/uploads/2018/12/btw2019-jmkeil-camera-ready.pdf>.
- [5] Leonard Kaufman, Peter J.Rousseeuw. *Finding Groups In Data: An Introduction to Cluster Analysis*.
- [6] Greg Hamerly, Charles Elkan. *Learning the k in k-means*. URL: <https://proceedings.neurips.cc/paper/2003/file/234833147b97bb6aed53a8f41Paper.pdf>.
- [7] Overflow, Stack. *2020 Developer Survey*. 2020. URL: <https://insights.stackoverflow.com/survey/2020#most-popular-technologies>.

- [8] Facebook. *The Open Graph protocol*. URL: <https://ogp.me>.
- [9] Mozilla. *Mozilla Readability*. URL: <https://github.com/mozilla/readability>.