
TRƯỜNG ĐẠI HỌC PHENIKAA
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO ĐỒ ÁN CƠ SỞ

LDA

Khai phá chủ đề và xu hướng của các bài báo khoa học sử dụng Topic Modeling

Nguyễn Quang Nhật – 22010510 – 22010510@st.phenikaa-uni.edu.vn

Bùi Quốc Việt - 22010047 - 22010047@st.phenikaa-uni.edu.vn

Hồ Xuân Hùng – 22010493 - 22010493@st.phenikaa-uni.edu.vn

Tiến sĩ Phạm Tiến Lâm

Hà Nội, 22/07/2004

Content

1. Giới thiệu	3
1.1 Đặt vấn đề.....	3
1.2 Các giải pháp đã có.....	3
1.3 Giải pháp đề xuất.....	4
1.3.1. Mô tả tổng quan về giải pháp đề xuất.	4
1.3.2 Phân tích xu hướng bằng mô hình LDA	4
2. Thiết kế và triển khai	5
2.1 Các yêu cầu chức năng.....	5
2.1.1 Phân tích chủ đề của văn bản	5
2.1.2. Công cụ tìm kiếm :	6
2.1.3: Khai phá xu hướng:.....	7
2.2 Các yêu cầu phi chức năng.....	9
2.2.1. Hiệu năng và Khả năng mở rộng.....	9
2.2.2. Tính khả dụng và Độ tin cậy	10
2.2.3. Khả năng sử dụng	10
2.2.4. Tính chính xác	10
2.3 Các ràng buộc (Constraints).....	11
2.4 Các ràng buộc về triển khai	12
2.4.1 Các ràng buộc kinh tế	12
2.4.2 Các ràng buộc về đạo đức.....	13
2.5 Mô hình hệ thống / Thiết kế giải pháp.....	14
2.5.1 Các Kịch bản Sử dụng.....	14
2.5.2 Mô hình Use-case.....	18
2.5.3 Các màn hình giao diện người dùng.....	19
3. Một số thành phần khác của đề án.....	31
3.1. Kế hoạch dự án	31
3.1.1. Model Training:.....	31
3.1.2. Article Search:	31
3.1.3. Trend Analysis:	31
3.1.4. Document Analysis:.....	32
3.1.5. Report Writing:	32
3.2. Kế hoạch cho kiến thức mới và chiến lược học tập.....	33
Nhóm đã thành công hoàn thành project, nhưng nhận thấy còn có nhiều thiếu sót trong việc lên kế hoạch và phân bổ công việc phù hợp, định hướng và các công việc cần làm chưa rõ ràng đã khiến dự án hoàn thành chậm hơn so với dự kiến. Một trong những kế hoạch học kiến thức mới là tìm hiểu về cách quản trị một dự án. Ngoài ra, chúng tôi còn nhận thấy sự rộng lớn của kiến thức Machine Learning và định hướng sẽ tìm hiểu thêm để cải thiện khả năng, kiến thức của mình trong lĩnh vực này.	
4. Kết luận	33
5. Tài liệu tham khảo	34

1. Giới thiệu

1.1 Đặt vấn đề

Trong kỷ nguyên thông tin hiện nay, số lượng bài báo khoa học xuất bản hàng năm tăng nhanh chóng, gây khó khăn cho việc nắm bắt các chủ đề nổi bật và xu hướng nghiên cứu. Đọc và phân tích thủ công khối lượng lớn văn bản này là không khả thi. Do đó, cần có phương pháp tự động trích xuất thông tin quan trọng từ dữ liệu văn bản này.

1.2 Các giải pháp đã có

Phân tích các giải pháp/hệ thống tương tự đã có. Nêu một số hạn chế của các giải pháp đó.

Một số phương pháp đã được sử dụng để giải quyết vấn đề này, bao gồm:

- Phân tích tần suất từ: Đếm số lần xuất hiện của từ, bỏ qua ngữ cảnh và ý nghĩa tiềm ẩn.
- Phân cụm văn bản: Nhóm các bài báo tương tự nhau, nhưng không xác định được chủ đề cụ thể.
- Phân tích ngữ nghĩa tiềm ẩn (LSA): Sử dụng kỹ thuật đại số tuyến tính, nhưng kết quả khó diễn giải.

Các phương pháp này có những hạn chế nhất định trong việc nắm bắt đầy đủ các khía cạnh của chủ đề và xu hướng trong các bài báo khoa học.

1.3 Giải pháp đề xuất

1.3.1. Mô tả tổng quan về giải pháp đề xuất.

Giải pháp đề xuất tập trung vào việc sử dụng Mô hình Latent Dirichlet Allocation (LDA), một kỹ thuật học máy không giám sát mạnh mẽ, để khám phá các chủ đề tiềm ẩn trong tập dữ liệu bài báo khoa học. LDA có khả năng tự động nhóm các từ có liên quan về mặt ngữ nghĩa thành các chủ đề, giúp chúng ta hiểu rõ hơn về nội dung của các bài báo.

LDA giả định rằng mỗi bài báo là một hỗn hợp của các chủ đề, và mỗi chủ đề là một phân bố xác suất trên các từ. Bằng cách phân tích phân bố từ trong mỗi chủ đề, chúng ta có thể hiểu được ý nghĩa và nội dung của chủ đề đó.

Ưu điểm của mô hình LDA:

Xác định chủ đề cụ thể: LDA có thể xác định các chủ đề cụ thể trong tập dữ liệu, giúp chúng ta hiểu rõ hơn về nội dung của các bài báo.

Theo dõi sự thay đổi của chủ đề theo thời gian: LDA có thể theo dõi sự thay đổi của các chủ đề qua thời gian, giúp chúng ta nắm bắt xu hướng nghiên cứu.

1.3.2 Phân tích xu hướng bằng mô hình LDA

Để phân tích xu hướng, chúng ta có thể áp dụng mô hình LDA trên các tập dữ liệu bài báo khoa học theo từng giai đoạn thời gian khác nhau. Sau đó, so sánh kết quả của các mô hình LDA này để tìm ra sự thay đổi của các chủ đề theo thời gian.

Cụ thể, ta có thể thực hiện các bước sau:

Chia tập dữ liệu: Chia tập dữ liệu bài báo khoa học thành các tập con theo từng giai đoạn thời gian (ví dụ: theo năm, theo quý).

Huấn luyện mô hình LDA: Huấn luyện mô hình LDA trên từng tập con dữ liệu.

Phân tích kết quả: Phân tích kết quả của các mô hình LDA để tìm ra:

- Các chủ đề mới xuất hiện: Các chủ đề chỉ xuất hiện ở những giai đoạn sau.
- Các chủ đề biến mất: Các chủ đề chỉ xuất hiện ở những giai đoạn trước.
- Các chủ đề thay đổi về mức độ phổ biến: Các chủ đề có sự thay đổi về tỷ lệ đóng góp trong tập dữ liệu theo thời gian.

Bằng cách phân tích sự thay đổi của các chủ đề theo thời gian, chúng ta có thể nhận diện được các xu hướng nghiên cứu mới nổi, các lĩnh vực đang được quan tâm nhiều hơn, và các chủ đề đang dần mất đi sự quan tâm.

Phân tích xu hướng nghiên cứu có thể giúp các nhà nghiên cứu, nhà hoạch định chính sách và doanh nghiệp đưa ra quyết định sáng suốt hơn. Ví dụ, các nhà nghiên cứu có thể xác định được các lĩnh vực nghiên cứu tiềm năng, các nhà hoạch định chính sách có thể đưa ra các chính sách hỗ trợ nghiên cứu phù hợp, và các doanh nghiệp có thể phát triển các sản phẩm và dịch vụ mới dựa trên xu hướng thị trường.

2. Thiết kế và triển khai

2.1 Các yêu cầu chức năng

2.1.1 Phân tích chủ đề của văn bản

Phân loại chủ đề:

Người dùng có thể cung cấp văn bản hoặc đường dẫn đến file chứa văn bản để khai phá chủ đề tiềm ẩn trong dữ liệu. Quy trình này sử dụng mô hình LDA đã được huấn luyện để dự đoán xác suất của văn bản đối với từng chủ đề dựa vào các trọng số tìm được trong

quá trình huấn luyện. Kết quả cuối cùng sẽ danh sách các chủ đề mà văn bản có xác suất cao nhất.

Topic Radar:

Trực quan hóa xác suất đó bằng đồ thị Radar, giúp người dùng dễ dàng nắm bắt hơn văn bản có phân phối xác suất thuộc các chủ đề như thế nào.

Word cloud:

Một trong những cách hiệu quả để trình bày kết quả phân loại chủ đề là thông qua word cloud. Hệ thống sẽ hiển thị word cloud của phân phối từ trong mô hình, với 20 từ có trọng số cao nhất trong mỗi chủ đề. Word cloud là một biểu đồ trực quan giúp người dùng dễ dàng nhận biết các từ khóa quan trọng nhất trong từng chủ đề. Những từ có trọng số cao hơn sẽ có kích thước lớn hơn trong word cloud, giúp người dùng nhanh chóng nắm bắt được nội dung chính của các chủ đề khác nhau.

2.1.2. Công cụ tìm kiếm :

Tìm kiếm các bài báo liên quan:

Hệ thống cho phép người dùng thực hiện tìm kiếm các bài báo liên quan bằng cách sử dụng mô hình dự đoán chủ đề để phân tích đầu vào tìm kiếm của người dùng. Quá trình này sẽ xác định chủ đề của từ khóa tìm kiếm và hiển thị các kết quả có cùng chủ đề với chủ đề mà mô hình đã dự đoán từ tìm kiếm của người dùng.

Kết quả tìm kiếm sẽ được hiển thị dưới dạng danh sách các bài báo, mỗi bài báo bao gồm các thông tin sau:

- **Correlation Confidence:** Xác suất mà mô hình dự đoán rằng bài báo thuộc cùng chủ đề được tìm kiếm.
- **Title:** Tiêu đề chính của bài báo.
- **ID:** Mã định danh duy nhất của bài báo trong tập dữ liệu
- **DOI:** Digital Object Identifier (DOI), giúp dễ dàng truy cập và tham khảo bài báo đầy đủ trên các trang chính thức
- **Abstract:** Phần tóm tắt nội dung chính của bài báo, giúp người dùng nhanh chóng đánh giá tính liên quan của bài báo với nhu cầu tìm kiếm của mình.

*Xác định chủ đề :

Hệ thống có khả năng nhận một văn bản từ người dùng và dự đoán chủ đề của văn bản đó. Quá trình này giúp người dùng hiểu rõ hơn về nội dung và các chủ đề chính mà văn bản đề cập.

Sau khi phân tích, hệ thống sẽ hiển thị một biểu đồ radar (radar chart) thể hiện xác suất của văn bản với mỗi chủ đề. Biểu đồ radar là một công cụ trực quan giúp người dùng dễ dàng so sánh và đánh giá mức độ liên quan của văn bản với các chủ đề khác nhau, từ đó đưa ra các quyết định phân loại và sử dụng thông tin một cách hiệu quả hơn.

2.1.3: Khai phá xu hướng:

Chức năng hiển thị xu hướng là một phần quan trọng trong hệ thống phân tích dữ liệu, giúp người dùng theo dõi và đánh giá các xu hướng của bài viết theo thời gian. Chức năng này dựa trên việc sử dụng mô hình LDA (Latent Dirichlet Allocation) để phân loại các chủ đề của bài viết và từ đó phân tích xu hướng theo thời gian dựa trên ngày đăng của các bài viết và chủ đề tìm được từ mô hình với mỗi bài.

a, Nhập thông tin khoảng thời gian

Chức năng hiển thị xu hướng cho phép người dùng nhập thông tin về khoảng thời gian muốn kiểm tra xu hướng. Các kết quả sẽ được hiển thị gộp theo ngày, tháng hoặc năm, số chủ đề muốn hiển thị, và đường link đến bảng thể hiện số bài viết thuộc chủ đề được thống kê theo ngày. Nếu thông tin không được cung cấp, hệ thống sẽ sử dụng các giá trị mặc định như sau:

- **Ngày bắt đầu:** Ngày sớm nhất trong dữ liệu.
- **Ngày kết thúc:** Ngày cuối cùng trong dữ liệu.
- **Số chủ đề:** Tổng số chủ đề có trong dữ liệu
- **Đường link:** Đường link đến bảng thông tin có sẵn của hệ thống

b, Hiển thị thông tin xu hướng

Sau khi nhận được các thông tin từ người dùng, hệ thống sẽ tiến hành hiển thị các biểu đồ và thông số thể hiện xu hướng của các bài viết. Các thông tin này bao gồm:

- **Biểu đồ tuyến tính:** Biểu đồ này sẽ hiển thị số lượng bài viết thuộc mỗi chủ đề theo thời gian. Người dùng có thể dễ dàng nhận ra sự thay đổi trong xu hướng của các chủ đề qua các khoảng thời gian khác nhau.
- **Biểu đồ Stacked Bar Chart:** Biểu đồ này sẽ hiển thị số lượng bài viết thuộc mỗi chủ đề theo thời gian dưới dạng các cột xếp chồng lên nhau. Điều này giúp người dùng có cái nhìn tổng quan về sự phân bố của các chủ đề trong từng khoảng thời gian.
- **Bảng ghi chú số bài viết:** Bảng này sẽ ghi chú số lượng bài viết thuộc mỗi chủ đề và hiển thị theo kiểu gộp mà người dùng đã lựa chọn (theo ngày, tháng, hoặc năm)

c, Xử lý thông tin thời gian

Trong trường hợp người dùng nhập thông tin thời gian không cụ thể về ngày hoặc tháng, hệ thống sẽ tự động xử lý như sau:

- **Thời điểm bắt đầu:**
 - Nếu chỉ nhập tháng mà không nhập ngày, hệ thống sẽ lấy ngày đầu tiên của tháng đó.
 - Nếu chỉ nhập năm mà không nhập tháng, hệ thống sẽ lấy tháng sớm nhất của năm đó.
- **Thời điểm kết thúc:**
 - Nếu chỉ nhập tháng mà không nhập ngày, hệ thống sẽ lấy ngày cuối cùng của tháng đó.

- Nếu chỉ nhập năm mà không nhập tháng, hệ thống sẽ lấy tháng cuối cùng của năm đó.

Nếu khoảng thời gian lựa chọn nằm ngoài phạm vi dữ liệu, hệ thống sẽ tự động chọn ngày gần nhất có trong dữ liệu để thay thế.

d, Phạm vi thông tin được hiển thị

Thông tin được tổng hợp và hiển thị sẽ chỉ nằm trong khoảng thời gian mà người dùng cung cấp. Điều này đảm bảo rằng người dùng chỉ nhận được những thông tin liên quan và phù hợp với nhu cầu của họ.

2.2 Các yêu cầu phi chức năng

Trong kỷ nguyên bùng nổ thông tin, dữ liệu khoa học đang tăng trưởng với tốc độ chóng mặt, tạo ra một kho tàng tri thức vô giá nhưng cũng đặt ra thách thức lớn trong việc khai thác và phân tích hiệu quả. Dự án "Phát hiện chủ đề và xu hướng trong nghiên cứu khoa học sử dụng topic modeling với thuật toán LDA" trên tập dữ liệu 1 triệu abstract là một nỗ lực đáng chú ý để giải quyết vấn đề này. Tuy nhiên, để dự án thành công, việc đáp ứng các yêu cầu phi chức năng (NFRs) là vô cùng quan trọng.

2.2.1. Hiệu năng và Khả năng mở rộng

Đối với tập dữ liệu 1 triệu abstract, yêu cầu về hiệu năng đòi hỏi hệ thống phải có khả năng xử lý khối lượng dữ liệu khổng lồ trong thời gian hợp lý. Mục tiêu là hoàn thành phân tích toàn bộ tập dữ liệu trong vòng 24 giờ và đảm bảo thời gian phản hồi nhanh chóng cho các truy vấn tìm kiếm xu hướng (dưới 10 giây). Việc tối ưu hóa thuật toán LDA, sử dụng các kỹ thuật tính toán phân tán

và cân nhắc phần cứng chuyên dụng như GPU là rất cần thiết. Hệ thống cần có khả năng mở rộng theo cả chiều dọc (nâng cấp phần cứng) và chiều ngang (mở rộng cụm máy chủ) để đáp ứng nhu cầu xử lý dữ liệu ngày càng tăng trong tương lai. Áp dụng các kỹ thuật lưu trữ thông minh như phân vùng dữ liệu, nén dữ liệu và lưu trữ phân tán sẽ giúp tối ưu hóa hiệu suất hệ thống một cách hiệu quả.

2.2.2. Tính khả dụng và Độ tin cậy

Tính khả dụng cao là yếu tố then chốt, đặc biệt đối với các hệ thống xử lý dữ liệu lớn. Yêu cầu hệ thống phải có thời gian hoạt động tối thiểu 99.9% trong một tháng, nghĩa là hệ thống chỉ được phép ngừng hoạt động không quá 43.2 phút mỗi tháng. Điều này đòi hỏi hệ thống phải có khả năng tự động phát hiện và khắc phục sự cố nhanh chóng, đồng thời có cơ chế dự phòng dữ liệu mạnh mẽ để đảm bảo dịch vụ không bị gián đoạn. Sử dụng các công nghệ như ảo hóa, container hóa và sao lưu dữ liệu tự động sẽ tăng cường tính khả dụng và độ tin cậy của hệ thống.

2.2.3. Khả năng sử dụng

Giao diện người dùng trực quan, thân thiện là yếu tố quan trọng để thu hút và duy trì người dùng. Tính năng tùy chỉnh, tìm kiếm và lọc nâng cao sẽ giúp người dùng dễ dàng khám phá và phân tích dữ liệu theo nhu cầu. Cá nhân hóa trải nghiệm người dùng thông qua việc lưu trữ các truy vấn và cài đặt tùy chỉnh sẽ tăng tính tiện dụng và hiệu quả của hệ thống.

2.2.4. Tính chính xác

Để đảm bảo tính chính xác của kết quả phân tích, hệ thống cần sử dụng nhiều phương pháp đánh giá mô hình LDA khác nhau. Người dùng cần có khả

năng đánh giá và cung cấp phản hồi về kết quả, từ đó hệ thống có thể điều chỉnh và cải thiện mô hình. Thường xuyên cập nhật mô hình với dữ liệu mới sẽ đảm bảo rằng kết quả phản ánh đúng các xu hướng và chủ đề nghiên cứu mới nhất.

2.3 Các ràng buộc (Constraints)

Dự án phát hiện chủ đề và xu hướng nghiên cứu khoa học với tập dữ liệu lớn mang lại những cơ hội lớn trong việc khai phá và phân tích dữ liệu khoa học. Tuy nhiên, để vượt qua những thách thức về nguồn lực, nhóm của chúng tôi đã phải đưa ra các giải pháp cụ thể và hiệu quả.

2.3.1. Thách thức về phần cứng:

Dự án của chúng tôi không có sẵn nguồn lực phần cứng mạnh mẽ để xử lý tập dữ liệu lớn. Thay vào đó, chúng tôi đã tận dụng các máy tính cá nhân và dịch vụ đám mây miễn phí/giá rẻ để giảm bớt áp lực về phần cứng. Điều này đã cho phép chúng tôi tối ưu hóa thuật toán LDA bằng cách sử dụng các thư viện tính toán trên CPU và tận dụng tối đa tài nguyên hiện có.

2.3.2. Thách thức về nhân sự:

Với chỉ ba thành viên, việc phân chia công việc rõ ràng và hiệu quả là điều cực kỳ quan trọng. Mỗi thành viên trong nhóm tập trung vào một lĩnh vực chuyên môn cụ thể và đồng thời hỗ trợ nhau để đảm bảo tiến độ dự án. Sự hợp tác chặt chẽ này không chỉ giúp chúng tôi nâng cao hiệu suất làm việc mà còn tăng tính sáng tạo và sự đổi mới trong quá trình nghiên cứu.

2.3.3. Thách thức về thời gian:

Với hạn chế về nguồn lực, việc đặt ra mục tiêu hoàn thành dự án trong thời gian ngắn có thể không khả thi. Thay vào đó, chúng tôi đã chia nhỏ dự án thành các giai đoạn nhỏ, tập trung vào các tính năng cốt lõi và đặt ra các mốc thời gian hợp lý cho từng giai đoạn. Điều này giúp chúng tôi duy trì sự linh hoạt trong quá trình phát triển dự án mà vẫn đảm bảo chất lượng và hiệu quả công việc.

2.3.4. Mục tiêu đánh giá chất lượng:

Chúng tôi xác định rõ các tiêu chí đánh giá như độ chính xác của mô hình và độ tin cậy của hệ thống. Mặc dù không thể đạt được hiệu suất tối ưu như các dự án có nguồn lực lớn, nhóm của chúng tôi đã hướng tới việc xây dựng một hệ thống hoạt động ổn định và cung cấp kết quả phân tích có giá trị.

2.4 Các ràng buộc về triển khai

2.4.1 Các ràng buộc kinh tế

- **Chi phí phần cứng:** Việc xử lý tập dữ liệu lớn đòi hỏi máy chủ mạnh mẽ, bộ nhớ lớn và có thể cả GPU. Tuy nhiên, chi phí đầu tư cho các thiết bị này có thể vượt quá khả năng của nhiều nhóm nghiên cứu.
- **Chi phí dịch vụ đám mây:** Mặc dù dịch vụ đám mây cung cấp khả năng mở rộng và linh hoạt, chi phí sử dụng có thể tăng cao nếu không được quản lý cẩn thận.
- **Chi phí bảo trì và vận hành:** Việc duy trì và cập nhật hệ thống cũng đòi hỏi chi phí đáng kể về nhân lực và tài chính.

Giải pháp cho Ràng buộc Kinh Tế

Để giải quyết vấn đề chi phí phần cứng, nhóm nghiên cứu có thể tận dụng các máy tính cá nhân hiện có hoặc sử dụng các dịch vụ đám mây miễn phí/giá rẻ như Google Colaboratory, google Driver, HuggingFace Space. Đối với chi phí phần mềm, việc ưu tiên sử dụng các công cụ mã nguồn mở như Python, và Gensim không chỉ tiết kiệm chi phí mà còn cho phép tùy chỉnh linh hoạt. Nhóm nghiên cứu cố gắng tận dụng tất cả các nguồn lực miễn phí, có chi phí rẻ nhiều nhất có thể để tích kiệm chi phí hoạt động của nhóm.

2.4.2 Các ràng buộc về đạo đức

- **Bảo vệ dữ liệu:** Dữ liệu nghiên cứu, đặc biệt là thông tin cá nhân của các tác giả, cần được bảo vệ một cách nghiêm ngặt. Việc tuân thủ các quy định về bảo vệ dữ liệu.
- **Đạo đức nghiên cứu:** Các nguyên tắc đạo đức nghiên cứu như minh bạch, công bằng và khách quan cần được tuân thủ trong quá trình thu thập, phân tích và công bố kết quả nghiên cứu.

Giải pháp cho Ràng buộc Đạo Đức

Nhóm nghiên cứu cần thiết lập các quy trình bảo vệ dữ liệu nghiêm ngặt, như giới hạn quyền truy cập và lưu trữ an toàn. Việc ẩn danh hóa dữ liệu và xin phép tác giả trước khi sử dụng cũng là những biện pháp cần thiết. Hơn nữa, việc đào tạo và nâng cao nhận thức về bảo vệ dữ liệu và đạo đức nghiên cứu cho các thành viên trong nhóm là rất quan trọng. Việc tham khảo ý kiến từ các chuyên gia pháp lý và đạo đức cũng sẽ giúp đảm bảo tính tuân thủ và trách nhiệm của dự án.

2.5 Mô hình hệ thống / Thiết kế giải pháp

Trong bối cảnh bùng nổ thông tin khoa học, việc nắm bắt và phân tích các xu hướng nghiên cứu trở nên cực kỳ quan trọng đối với cộng đồng học thuật và các bên liên quan. Nhằm giải quyết thách thức này, một hệ thống thông minh được thiết kế để tự động phát hiện chủ đề và xu hướng nghiên cứu, sử dụng mô hình Topic Modeling với LDA (Latent Dirichlet Allocation) làm nền tảng. Hệ thống này hứa hẹn mang lại những giá trị to lớn cho việc nghiên cứu và phát triển khoa học, đồng thời cung cấp cái nhìn toàn diện và sâu sắc về bức tranh nghiên cứu hiện tại.

2.5.1 Các Kịch bản Sử dụng

2.5.1.1 Lựa chọn công cụ muốn sử dụng:

Người sử dụng : Người dùng

Điều kiện bắt đầu: Người dùng mở ứng dụng hoặc hoàn thành trả về kết quả

Điều kiện thoát: Người dùng đóng ứng dụng

Lưu đồ sự kiện :

- Người dùng chọn loại công cụ muốn sử dụng (“Trend Analysis”, “Article Search”, “Document Analysis”)
- Hệ thống hiển thị những thông tin đầu vào cần thiết tương ứng với loại công cụ được lựa chọn

2.5.1.2. Tìm kiếm và Hiển thị Bài báo theo Từ khóa:

Cho phép người dùng tìm kiếm các bài báo khoa học liên quan đến một lĩnh vực cụ thể bằng cách nhập từ khóa. Hệ thống sẽ xử lý từ khóa, truy vấn cơ sở dữ

liệu và trả về danh sách các bài báo phù hợp, giúp người dùng tiết kiệm thời gian và công sức nghiên cứu.

Người sử dụng: Người dùng

Điều kiện bắt đầu: Người dùng nhấn “Search” trong công cụ “Article Search”

Điều kiện kết thúc: Người dùng nhận được danh sách các bài báo phù hợp hoặc đóng ứng dụng

Lưu đồ sự kiện:

1. Nhận thông tin:

- Hệ thống xử lý truy vấn người dùng nhập vào.
- Nếu người dùng để trống / hoặc thông tin nhập vào chỉ gồm khoảng trắng, hệ thống sẽ yêu cầu người dùng nhập truy vấn.

2. Xử lý thông tin :

- Truy vấn được xử lý bằng mô hình LDA đã qua huấn luyện và trả về các ID của các chủ đề có xác suất cao nhất.
- Truy vấn từ ID của chủ đề trong cơ sở dữ liệu và trả về các bài báo thuộc chủ đề đó.

3. Hiển thị kết quả

- Hệ thống hiển thị danh sách các bài báo từ kết quả truy vấn cơ sở dữ liệu
- Nếu không tìm, hệ thống thông báo không tìm được kết quả nào.
- Mỗi mục trong danh sách sẽ bao gồm: xác suất bài báo thuộc cùng chủ đề với truy vấn của người dùng, tiêu đề bài báo, DOT của bài báo.
- Người dùng có thể nhấp vào từng mục để xem chi tiết tóm tắt của bài báo.

2.5.1.3. Phân tích chủ đề của văn bản:

Hệ thống phân tích chủ đề, cho người dùng biết xác suất chủ đề của văn bản thông qua dự đoán của mô hình.

Người sử dụng: Người dùng

Điều kiện bắt đầu: Người dùng nhấn “Analyze” trong công cụ “Document Analysis”

Điều kiện kết thúc: Người dùng nhận được kết quả phân tích hoặc đóng ứng dụng

Lưu đồ sự kiện:

1. Nhận dữ liệu:

- Người dùng nhập văn bản hoặc nhập đường dẫn đến file văn bản

2. Xử lý thông tin:

- Nếu thông tin người dùng nhập vào là một đường dẫn, hệ thống sẽ mở file đó và đọc thông tin trong file, nếu không phải đường dẫn, hệ thống sẽ coi thông tin nhập vào là văn bản cần phân tích
- Sử dụng mô hình LDA, hệ thống sẽ xác định các chủ đề chính của văn bản.

3. Hiển thị kết quả:

- Hệ thống hiển thị các chủ đề có xác suất chủ đề có khả năng cao nhất theo thứ tự từ cao xuống thấp.
- Hiển thị biểu đồ Radar Chủ đề của văn bản.
- Hiển thị Word Cloud thể hiện 20 từ có trọng số cao nhất của chủ đề mà mô hình tìm ra trong quá trình huấn luyện.

2.5.1.4. Phân tích Xu hướng Nghiên cứu theo Thời gian:

Hệ thống cho phép người dùng theo dõi sự phát triển của các chủ đề nghiên cứu theo thời gian. Bằng cách phân tích dữ liệu theo các giai đoạn khác nhau, hệ thống cung cấp cái nhìn sâu sắc về sự thay đổi và phát triển của các lĩnh vực nghiên cứu.

Người sử dụng: Người dùng

Điều kiện bắt đầu: Người dùng nhấn “Show Trend” trong công cụ “Trend Analysis”

Điều kiện kết thúc: Người dùng nhận được phân tích xu hướng hoặc đóng ứng dụng

Lưu đồ sự kiện:

1. Nhận thông tin

- Người dùng cung cấp khoảng thời gian mong muốn.
- Lựa chọn thông tin hiển thị gộp theo ngày, tháng hoặc năm

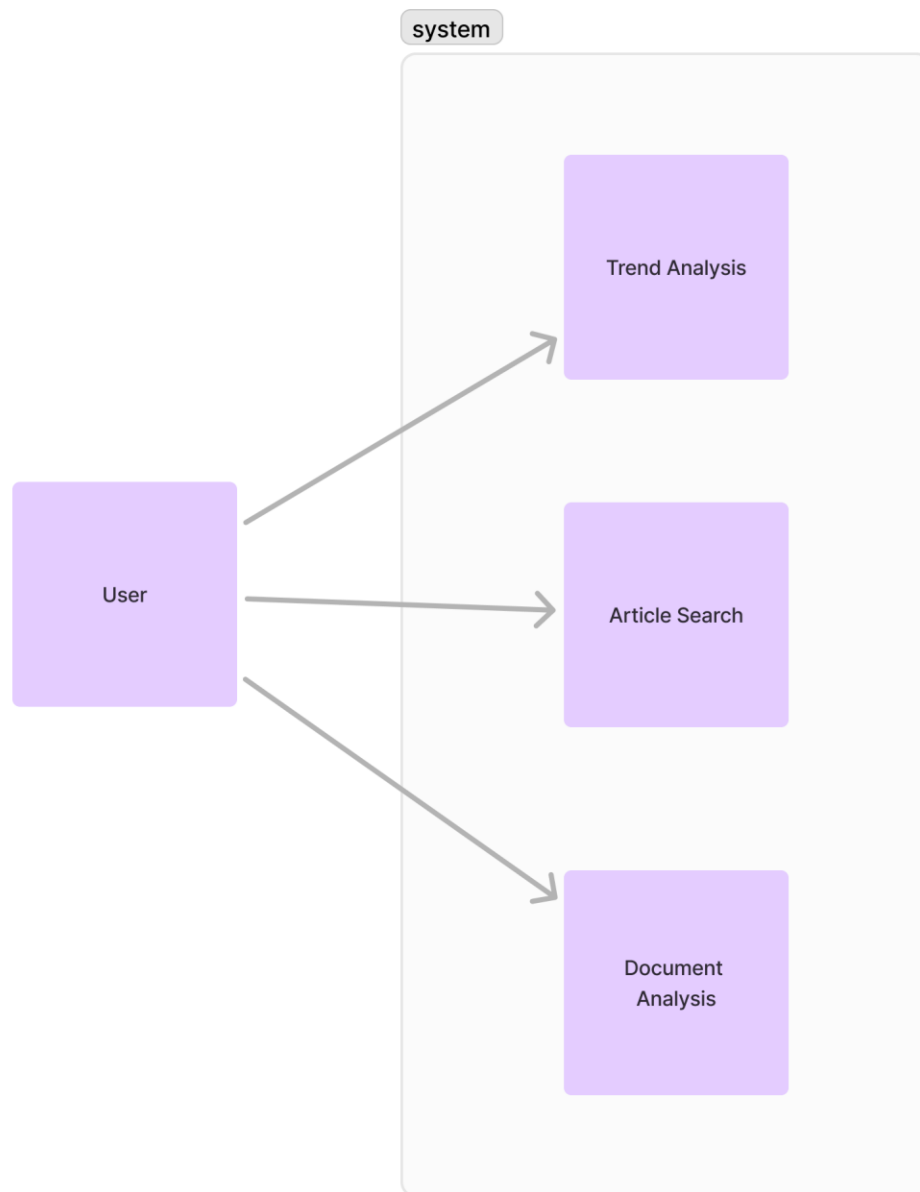
2. Xử lý thông tin:

- Hệ thống sẽ lấy đến ngày sớm nhất trong khoảng thời gian tìm kiếm và ngày cuối cùng của trong khoảng thời gian tìm kiếm
- Hệ thống phân tích dữ liệu bài báo trong khoảng thời gian đã chọn để xác định các xu hướng nghiên cứu.
- Nếu khoảng thời gian nằm ngoài bộ dữ liệu mà hệ thống sở hữu, khoảng thời gian sẽ được thu hẹp về khoảng thời gian nằm trong bộ dữ liệu của hệ thống.

3. Hệ thống hiển thị xu hướng nghiên cứu:

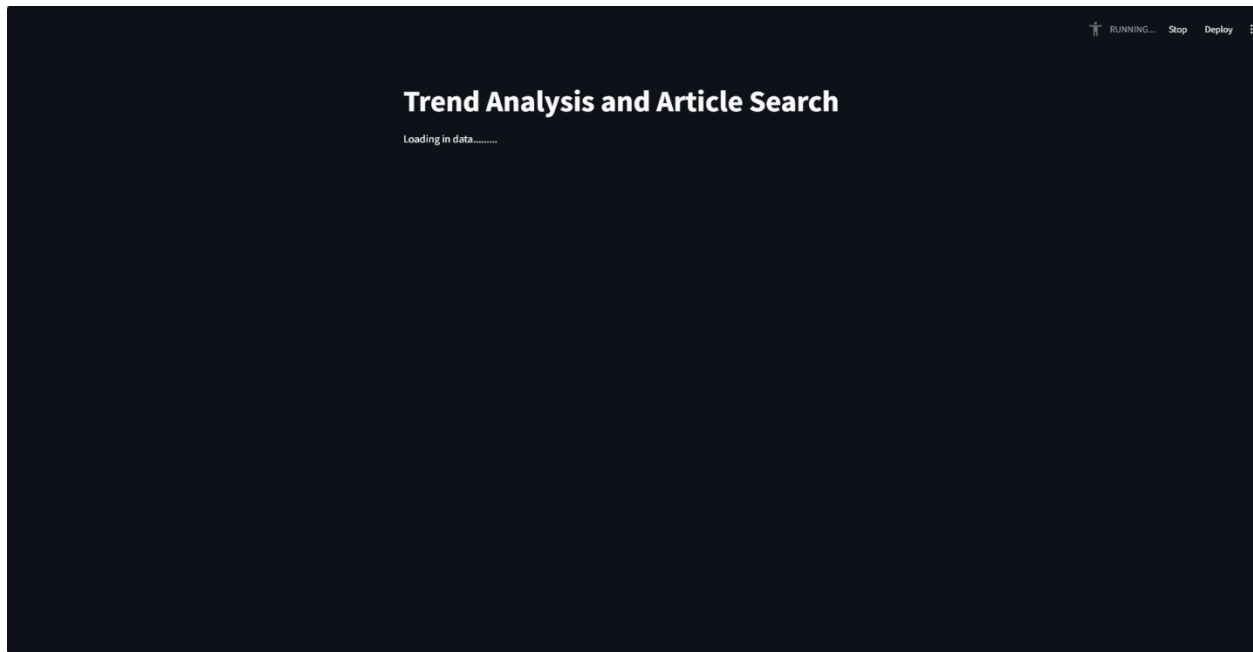
- Hệ thống hiển thị các biểu đồ hoặc báo cáo về xu hướng nghiên cứu trong khoảng thời gian đã chọn.
- Các biểu đồ này có thể bao gồm số lượng bài báo theo từng khoảng trong khoảng thời gian cung cấp và chủ đề.

2.5.2 Mô hình Use-case

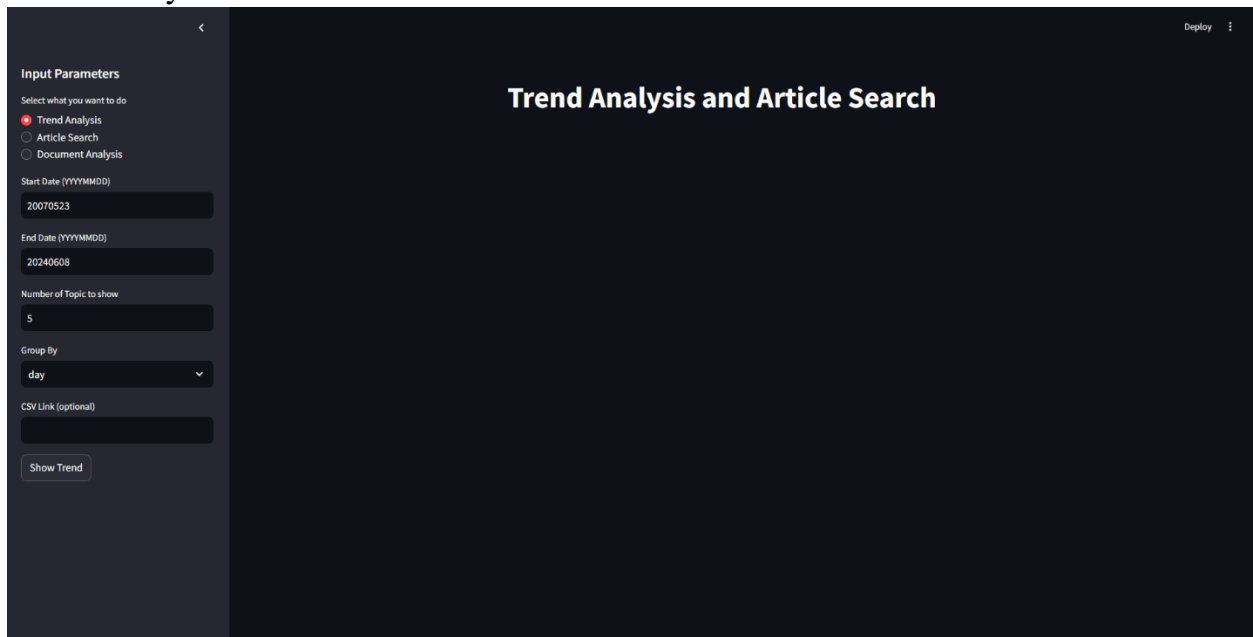


2.5.3 Các màn hình giao diện người dùng

Loading Screen



Trend Analysis Screen



Trend Analysis Result Screen

Input Parameters

Select what you want to do

- ☒ Trend Analysis
- ☐ Article Search
- ☐ Document Analysis

Start Date (YYYYMMDD)

20170523

End Date (YYYYMMDD)

20240608

Number of Topic to show

5

Group By

year

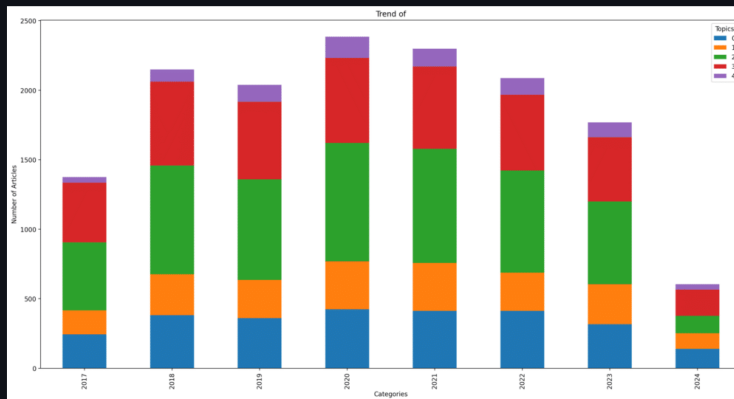
CSV Link (optional)

Show Trend

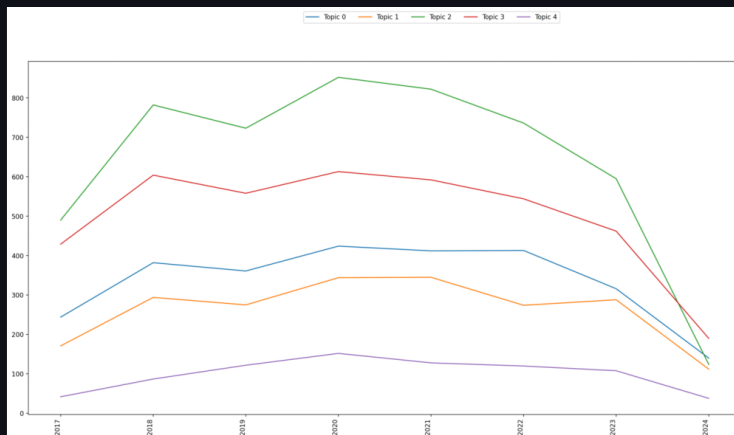
Trend Analysis and Article Search

	2017	2018	2019	2020	2021	2022	2023	2024
0	244	382	361	424	412	413	316	140
1	171	294	275	344	345	274	288	112
2	490	782	723	852	822	736	595	124
3	429	604	558	613	592	544	462	190
4	42	87	122	152	128	120	108	38
5	116	159	146	174	143	158	139	48
6	1,176	1,330	1,157	1,212	1,199	1,097	1,003	290
7	190	294	270	269	275	247	226	80
8	1,129	1,839	1,870	2,106	1,994	1,740	1,548	618
9	1	4	1	2	1	3	2	3

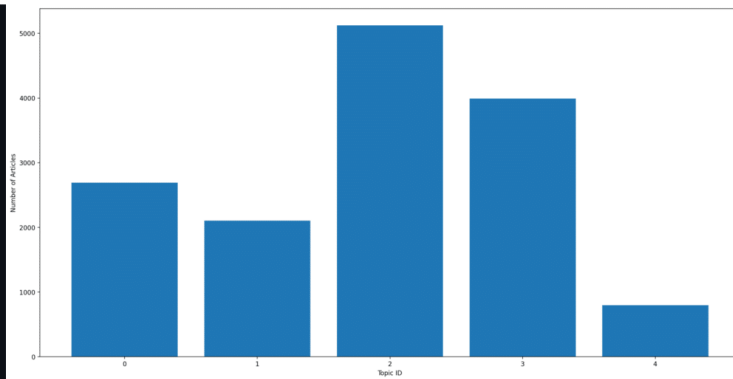
Bar Visualization



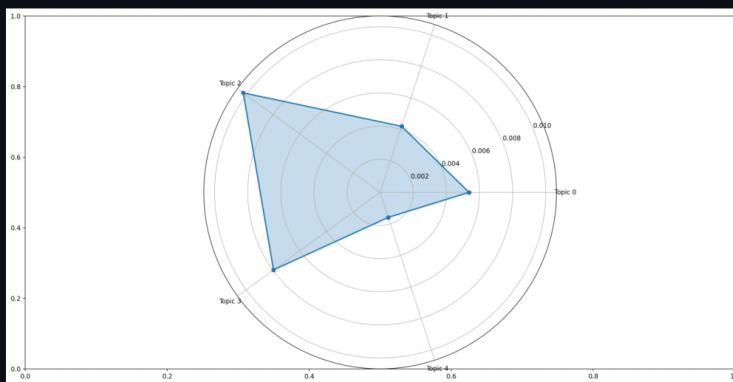
Line Visualization



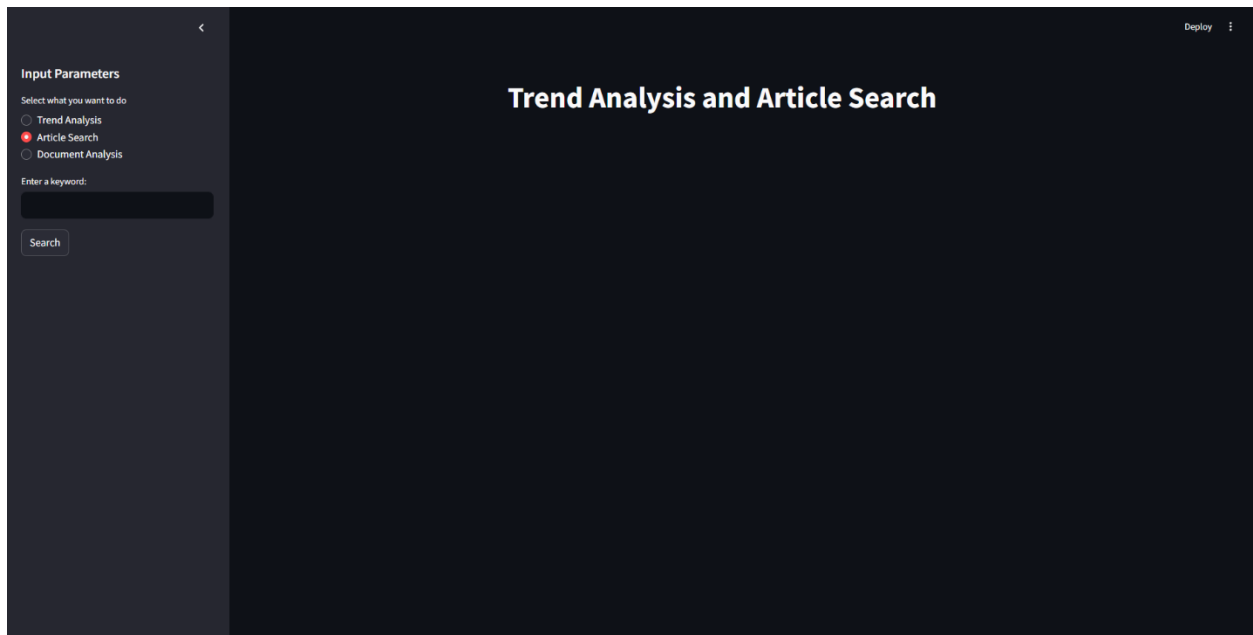
Total topics from 20170523 to 20240608



Rader Visualize of topics from 20170523 to 20240608



Article Search Screen



The screenshot shows a web application interface for "Trend Analysis and Article Search". On the left, there is a sidebar with the title "Input Parameters" and a back arrow. Below the title, it says "Select what you want to do" and lists three radio button options: "Trend Analysis", "Article Search" (which is selected and has a red dot), and "Document Analysis". Underneath, it says "Enter a keyword:" followed by a text input field and a "Search" button. The main area of the screen is dark blue and contains the title "Trend Analysis and Article Search" in white text. In the top right corner, there is a "Deploy" button and a menu icon.

Article Search Result Screen

Input Parameters

Select what you want to do

- ☐ Trend Analysis
- ☒ Article Search
- ☐ Document Analysis

Enter a keyword:

investing

Search

Trend Analysis and Article Search

Results for 'investing':

Correlation Confidence: 36.0%

ID: 2209.12222

Title: Efficient Wrong-Way Risk Modelling for Funding Valuation Adjustments

DOI: 10.1142/S0219024924500109

Abstract: Wrong-Way Risk (WWR) is an important component in Funding Valuation Adjustment (FVA) modelling. Yet, the standard assumption is independence between market risks and the counterparty defaults and fu...

Full Abstract

Correlation Confidence: 27.1%

ID: 2406.035

Title: Impact of aleatoric, stochastic and epistemic uncertainties on project cost contingency reserves

DOI: 10.1016/j.ijpe.2022.108626

Abstract: In construction projects, contingency reserves have traditionally been estimated based on a percentage of the total project cost, which is arbitrary and, thus, unreliable in practical cases. Monte C...

Full Abstract

Correlation Confidence: 41.4%

ID: 2401.02445

Title: Social and Economic Impact Analysis of Solar Mini-Grids in Rural Africa: A Cohort Study from Kenya and Nigeria

DOI: 10.1088/2634-4505/ad4ffb

Abstract: This study presents the first comprehensive analysis of the social and economic effects of solar mini-grids in rural African settings, specifically in Kenya and Nigeria. A group of 2,658 household h...

Full Abstract

Correlation Confidence: 27.5%

ID: 2406.02934

Title: Estimating Disease-Free Life Expectancy based on Clinical Data from the French Hospital Discharge Database

DOI: 10.3390/risks12060092

Abstract: The development of health indicators to measure healthy life expectancy (HLE) is an active field of research aimed at summarizing the health of a population. Although many health indicators have emerged...

Full Abstract



Correlation Confidence: 52.6%

ID: 2406.02297

Title: Optimal Stock Portfolio Selection with a Multivariate Hidden Markov Model

DOI: 10.1007/s13571-022-00290-5

Abstract: The underlying market trends that drive stock price fluctuations are often referred to in terms of bull and bear markets. Optimal stock portfolio selection methods need to take into account these ma...

Full Abstract



Correlation Confidence: 36.8%

ID: 2406.00056

Title: Optimizing Bio-energy Supply Chain to Achieve Alternative Energy Targets

DOI: 10.52783/jes.3176

Abstract: In response to global warming and the dwindling reservoirs of fossil fuels, Thailand has increasingly embraced alternative energy sources. Central to its energy development strategy is the Alternati...

Full Abstract



Correlation Confidence: 31.2%

ID: 2406.00078

Title: Project Risk Management from the bottom-up: Activity Risk Index

DOI: 10.1007/s10100-020-00703-8

Abstract: Project managers need to manage risks throughout the project lifecycle and, thus, need to know how changes in activity durations influence project duration and risk. We propose a new indicator (the ...

Full Abstract



Correlation Confidence: 48.0%

ID: 2406.00478

Title: Green Supply Chain Management Optimization Based on Chemical Industrial Clusters

DOI: 10.62836/iaet.v1i1.003

Abstract: Post-pandemic, the chemical sector faces new challenges crucial to national progress, with a pressing need for rapid transformation and upgrading. The pandemic's impact and increasing demand for sus...

Full Abstract



Correlation Confidence: 21.1%

ID: 2406.01399

Title: Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability

DOI: 10.1145/3630106.3658998

Abstract: In July 2023, New York City became the first jurisdiction globally to mandate bias audits for commercial algorithmic systems, specifically for automated employment decisions systems (AEDTs) used in ...

Full Abstract



Correlation Confidence: 21.9%

ID: 2405.16875

Title: Digitalization in Infrastructure Construction Projects: A PRISMA-Based Review of Benefits and Obstacles

DOI: 10.48550/arXiv.2405.16875

Abstract: The current study presents a comprehensive review of the benefits and barriers associated with the adoption of Building Information Modeling (BIM) in infrastructure projects, focusing on the period ...

Full Abstract



Document Analysis Screen

The screenshot shows a web application interface for document analysis. On the left is a dark sidebar with the title "Input Parameters" and a back arrow. It contains three radio buttons: "Trend Analysis", "Article Search", and "Document Analysis" (which is selected and has a red dot). Below these are three input fields: "Enter a Document or Document's Path:" with the value "C:/Users/japan/OneDrive/Desktop/super", "Number of top relevant topics showing:" with the value "5", and "LdaModel Address (Optional):" which is empty. At the bottom of the sidebar is an "Analyze" button. The main area on the right is dark blue and features the title "Trend Analysis and Article Search" in white. In the top right corner of the main area, there is a "Deploy" button and a menu icon.

Document Analysis Result Screen

Input Parameters

Select what you want to do

- ☐ Trend Analysis
- ☐ Article Search
- ☒ Document Analysis

Enter a Document or Document's Path:

C:/Users\japan\OneDrive\Desktop\super

Number of top relevant topics showing:

5

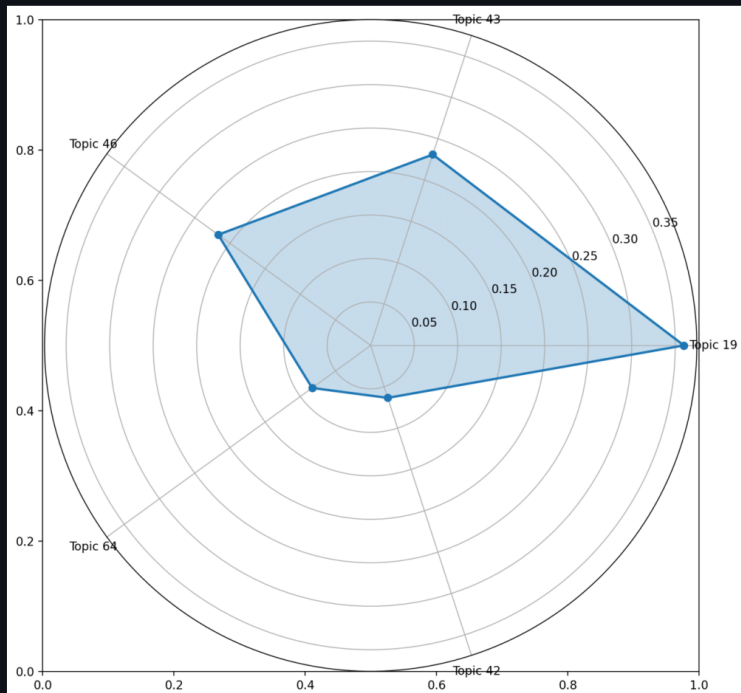
LdaModel Address (Optional):

Analyze

Trend Analysis and Article Search

	Topic ID	percent
0	19	36.0%
1	43	23.09%
2	46	21.66%
3	64	8.32%
4	42	6.34%

Radar Chart for the Document' topics



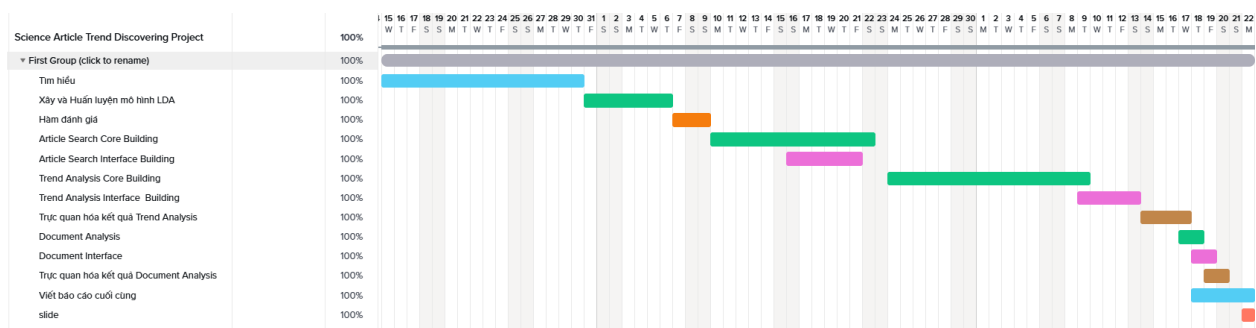
Word Cloud of Topic ID: 19

Topic 19



3. Một số thành phần khác của đề án

- Định hướng các công việc cần làm (Nhật): Mục tiêu của dự án là gì, ta giải quyết vấn đề gì, làm sao để đạt được mục tiêu, các milestone cần đạt, các công việc, nhiệm vụ cần thực hiện để đạt được mục tiêu dự án.
- Tìm hiểu về Topic Modeling (Hùng) : Topic Modeling là gì, sử dụng phương thức gì để giải quyết bài toán Topic Modeling, Phương thức đó hoạt động như thế nào
- Tìm hiểu cách để xây dựng interface cho dự án (Việt): Tìm hiểu các công cụ để xây dựng interface cho dự án, sử dụng công cụ đó như thế nào và xây dựng khuôn mẫu interface cho dự án.



3.1. Kế hoạch dự án

3.1.1. Model Training:

- Xử lý đầu vào (Việt) : tìm và lấy dữ liệu (sử dụng Arxiv API và Kaggle), tiền xử lý dữ liệu (loại bỏ stopwords, lemmatize, tokenize)
- Tính toán kết quả (Nhật): Sử dụng thư viện Gensim để xây dựng mô hình huấn luyện thực hiện Topic Modeling
- Chạy mô hình (Việt, Nhật) : chạy huấn luyện mô hình với số lượng chủ đề là : 3,5,10,20 (Việt), chạy huấn luyện mô hình với số lượng chủ đề là: 80 (Nhật)

3.1.2. Article Search:

- Xử lý đầu vào (Nhật) : Chuẩn bị bảng Topicid_to_ids để tra cứu bài báo theo topic
- Tính toán kết quả (Nhật): dự đoán truy vấn của người dùng thuộc topic nào và trả về các kết quả theo bảng Topicid_to_ids
- Hiển thị kết quả (Việt): sử dụng streamlit để tạo interface để tương tác với người dùng

3.1.3. Trend Analysis:

- Xử lý đầu vào (Nhật): đảm bảo dữ liệu đầu vào đúng kiểu, đưa dữ liệu đầu vào về đúng dạng chuẩn cho hệ thống, chuẩn bị bảng topic_trend_by_day để tra cứu, thống kê, tìm hiểu xu hướng các bài báo trong tập dữ liệu
- Tính toán kết quả (Nhật): tạo các hàm để lấy dữ liệu đúng trong các khoảng thời gian theo ngày, tháng, năm.
- Hiển thị kết quả (Việt, Nhật) : sử dụng streamlit để tạo interface để tương tác với người dùng (Việt), tạo hàm trực quan hóa các kết quả (Nhật)

3.1.4. Document Analysis:

- Xử lý đầu vào (Nhật): Đầu vào là một đường dẫn hay là văn bản, nếu là đường dẫn, mở file và đọc dữ liệu theo đường dẫn.
- Tính toán kết quả (Nhật): sử dụng mô hình đã qua huấn luyện để dự đoán văn bản đầu vào và trả về kết quả.
- Hiển thị kết quả (Hùng, Nhật) : tạo hàm trực quan hóa các kết quả (Hùng), sử dụng streamlit để tạo interface để tương tác với người dùng (Nhật)

3.1.5. Report Writing:

- Weekly report (Việt): Viết các bản báo cáo cho các buổi họp hằng tuần
- Final Report (Nhật, Việt, Hùng) :
 - + Phần 1,4 : Hùng
 - + Phần 2 : Việt, Nhật
 - + Phần 3: Nhật
- Slide (Hùng): làm slide cho bài thuyết trình bảo vệ đồ án cơ sở.

3.2. Kế hoạch cho kiến thức mới và chiến lược học tập

Nhóm đã thành công hoàn thành project, nhưng nhận thấy còn có nhiều thiếu sót trong việc lên kế hoạch và phân bổ công việc phù hợp, định hướng và các công việc cần làm chưa rõ ràng đã khiến dự án hoàn thành chậm hơn so với dự kiến. Một trong những kế hoạch học kiến thức mới là tìm hiểu về cách quản trị một dự án. Ngoài ra, chúng tôi còn nhận thấy sự rộng lớn của kiến thức Machine Learning và định hướng sẽ tìm hiểu thêm để cải thiện khả năng, kiến thức của mình trong lĩnh vực này.

4. Kết luận

Nêu một số kết luận của dự án. Nhóm đã thực hiện và hoàn thành được các nội dung gì. Tổng quan và đánh giá về kết quả đạt được.

- Sau bài báo cáo ta thấy Đồ án "LDA: Khai phá chủ đề và xu hướng của các bài báo khoa học sử dụng Topic Modeling" đã đạt được những kết quả tích cực trong việc xây dựng một hệ thống hoàn chỉnh, đáp ứng các yêu cầu chức năng đề ra. Cụ thể, hệ thống đã triển khai thành công các công cụ hỗ trợ phân tích văn bản, tìm kiếm thông tin và trực quan hóa dữ liệu, giúp người dùng dễ dàng khám phá và hiểu sâu hơn về các chủ đề và xu hướng nghiên cứu khoa học.
- Tuy nhiên, trong quá trình thực hiện, nhóm nghiên cứu đã phải đối mặt với một số hạn chế, đặc biệt trong việc xử lý và tối ưu hóa mô hình. Việc xử lý tập dữ liệu lớn đòi hỏi nhiều tài nguyên tính toán, trong khi nguồn lực phần cứng của nhóm còn hạn chế. Việc tối ưu hóa mô hình LDA để đạt hiệu suất tốt nhất cũng là một thách thức lớn, đòi hỏi sự tìm tòi và thử nghiệm nhiều giải pháp khác nhau. Bên cạnh đó, việc cải thiện giao diện người dùng để tăng tính thân thiện và dễ sử dụng cũng là một nhiệm vụ quan trọng cần được quan tâm trong tương lai.
- Mặc dù còn tồn tại những hạn chế nhất định, đồ án này đã khẳng định được tiềm năng của Topic Modeling trong việc khai phá tri thức từ dữ liệu văn bản

khoa học. Kết quả đạt được không chỉ đóng góp vào việc phát triển các công cụ hỗ trợ nghiên cứu mà còn mở ra hướng đi mới cho các ứng dụng thực tế khác trong lĩnh vực xử lý ngôn ngữ tự nhiên và khai phá dữ liệu. Nhóm nghiên cứu tự tin rằng với những nỗ lực cải tiến và hoàn thiện trong tương lai, hệ thống sẽ trở thành một công cụ hữu ích, đóng góp tích cực vào việc thúc đẩy sự phát triển của khoa học và công nghệ.

5. Tài liệu tham khảo

- [1] www.sciencedirect.com/science/article/abs/pii/S0968090X17300207
- [2] arxiv.org/pdf/1706.03762