ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN**

**NĂM 2023**

**PHÂN TÍCH DỮ LIỆU ĐỂ ĐÁNH GIÁ SỰ QUAN TÂM CỦA KHÁCH DU LỊCH VỀ CÁC ĐỊA ĐIỂM THAM QUAN LỊCH SỬ Ở VIỆT NAM**

**1**

**Lĩnh vực khoa học: Quản lý-Tin học**

**Chuyên ngành: Công nghệ thông tin**

**Nhóm nghiên cứu:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TT** | **Họ tên** | **MSSV** | **Đơn vị** | **Nhiệm vụ** | **Điện thoại** | **Email** |
| 1. | Nguyễn Tường Vy | K204162009 | Khoa Hệ thống thông tin | Nhóm trưởng | 0764187037 | vynt20416c@st.uel.edu.vn |
| 2. | Nguyễn Thị Xuân Trinh | K204131863 | Khoa Toán kinh tế | Tham gia | 0777663803 | trinhntx20413@st.uel.edu.vn |



ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN**

**NĂM 2023**

**Tên đề tài:**

**PHÂN TÍCH DỮ LIỆU ĐỂ ĐÁNH GIÁ SỰ QUAN TÂM CỦA KHÁCH DU LỊCH VỀ CÁC ĐỊA ĐIỂM THAM QUAN LỊCH SỬ Ở VIỆT NAM**

|  |  |  |
| --- | --- | --- |
| **Đại diện nhóm nghiên cứu**  **D:\Downloads\321020727_5579551052143158_4239091790992286827_n-removebg-preview.png**  **Nguyễn Tường Vy** | **Giảng viên hướng dẫn**  **D:\Downloads\unnamed.png**  **Nguyễn Thôn Dã** | **Chủ tịch Hội đồng** |

**Lãnh đạo Khoa/Bộ môn/Trung tâm**

*(Ký, họ tên)*

**TÓM TẮT ĐỀ TÀI**

Đề tài “Phân tích dữ liệu để đánh giá sự quan tâm của khách du lịch về các địa điểm tham quan lịch sử ở Việt Nam” nhằm nghiên cứu, phân tích và đánh giá thái độ của khách du lịch đến các địa điểm tham quan lịch sử ở Việt Nam thông qua việc phân tích dữ liệu các điểm đến văn hóa lịch sử tại Việt Nam trên trang Tripadvisor.com.

Đề tài này sẽ tập trung vào việc phân tích các yếu tố ảnh hưởng đến quan tâm của khách du lịch, bao gồm cả nhận xét, quan điểm, góc nhìn của du khách và các yếu tố tích cực, tiêu cực. Bằng việc phân tích trực quan tập dữ liệu và kết hợp áp dụng nhiều phương pháp xử lý ngôn ngữ khác nhau mà kết quả của đề tài sẽ là tư liệu tham khảo giúp các doanh nghiệp du lịch, cơ quan quản lý du lịch và các nhà nghiên cứu hiểu rõ hơn về thái độ của khách du lịch đến các địa điểm tham quan lịch sử ở Việt Nam, từ đó có thể đưa ra các chiến lược phát triển du lịch hiệu quả hơn.

**MỤC LỤC**

[**A. MỞ ĐẦU** 1](#_Toc129376398)

[**1. Giới thiệu** 1](#_Toc129376399)

[**3. Tổng quan nghiên cứu** 2](#_Toc129376400)

[**4. Mục tiêu nghiên cứu** 3](#_Toc129376401)

[**5. Đối tượng nghiên cứu** 4](#_Toc129376402)

[**6. Phạm vi nghiên cứu** 4](#_Toc129376403)

[**7. Phương pháp nghiên cứu** 4](#_Toc129376404)

[**B. NỘI DUNG NGHIÊN CỨU** 6](#_Toc129376405)

[**Chương 1: Cơ sở lý luận**: 6](#_Toc129376406)

[**Chương 2: Phân tích dữ liệu:** 13](#_Toc129376407)

[2.1. Mô tả dữ liệu 13](#_Toc129376408)

[2.2. Trực quan hóa dữ liệu 14](#_Toc129376409)

[2.3. Tạo WordCloud cho 3 tỉnh phổ biến nhất 19](#_Toc129376410)

[**Chương 3: Phân tích xúc theo khía cạnh** 21](#_Toc129376411)

[**C. KẾT LUẬN VÀ KIẾN NGHỊ** 27](#_Toc129376412)

[**1. Kết luận** 27](#_Toc129376413)

[**2. Đề nghị** 27](#_Toc129376414)

[**D. TÀI LIỆU THAM KHẢO** 28](#_Toc129376415)

DANH SÁCH BẢNG

[***Bảng 1: Data1 gồm 162 địa điểm về du lịch liên quan đến lịch sử 12***](#_Toc129165106)

[***Bảng 2: Data2 dữ liệu đánh giá, bình luận người dùng gồm 62 146 dòng 12***](#_Toc129165107)

DANH SÁCH HÌNH

[***Hình 1. Biểu đồ bong bóng thể hiện tổng số lượt đánh 13***](#_Toc129167281)

[***Hình 2. Biểu đồ map thể hiện phân phối số lượng địa điểm tại từng tỉnh 13***](#_Toc129167282)

[***Hình 3. Biểu đồ thể hiện cơ cấu số đánh giá 14***](#_Toc129167283)

[***Hình 4. Trung bình đánh giá ở mỗi tỉnh 14***](#_Toc129167284)

[***Hình 5. Số lượng bài viết qua các năm của từng ngôn ngữ 15***](#_Toc129167285)

[***Hình 6 . Loại hình du lịch theo từng ngôn ngữ 15***](#_Toc129167286)

[***Hình 7. Biểu đồ bản đồ thể hiện quốc tịch của người dùng 16***](#_Toc129167287)

[***Hình 8. Biểu đồ đường thể hiện xu hướng thời gian tham quan ở 2 ngôn ngữ 17***](#_Toc129167288)

[***Hình 9. WordCloud các thể loại của địa đến lịch sử 18***](#_Toc129167289)

[***Hình 10. WordCloud bình luận tiếng Việt và tiếng Anh ở thành phố Hồ Chí Minh. 18***](#_Toc129167290)

[***Hình 11. WordCloud bình luận tiếng Việt và tiếng Anh ở thủ đô Hà Nội 19***](#_Toc129167291)

[***Hình 12. WordCloud bình luận tiếng Việt và tiếng Anh ở tỉnh Quảng Nam 20***](#_Toc129167292)

[***Hình 13. 30 từ phổ biến nhất ở chủ đề 1 22***](#_Toc129167293)

[***Hình 14. 30 từ phổ biến nhất ở chủ đề 2 23***](#_Toc129167294)

[***Hình 15. 30 từ phổ biến nhất ở chủ đề 3 23***](#_Toc129167295)

[***Hình 16. Phân phối điểm số cảm xúc ở mỗi từ 24***](#_Toc129167296)

[***Hình 17. Word Cloud mô tả chủ đề 1 24***](#_Toc129167297)

[***Hình 18. Word Cloud mô tả chủ đề 2 24***](#_Toc129167298)

[***Hình 19. Word cloud mô tả chủ đề 3 24***](#_Toc129167299)

[***Hình 20. Biểu đồ Valence - Salience 25***](#_Toc129167300)

**PHÂN TÍCH DỮ LIỆU ĐỂ ĐÁNH GIÁ SỰ QUAN TÂM CỦA KHÁCH DU LỊCH VỀ CÁC ĐỊA ĐIỂM THAM QUAN LỊCH SỬ Ở VIỆT NAM**

# **A. MỞ ĐẦU**

## **1. Giới thiệu**

Việt Nam là một đất nước có rất nhiều địa điểm lịch sử và văn hóa hấp dẫn để khám phá. Với hơn 4.000 năm lịch sử văn hóa, Việt Nam đã có nhiều triều đại và giai đoạn lịch sử khác nhau, mỗi giai đoạn đều để lại những dấu ấn đặc biệt.

Vì thế mà Việt Nam là điểm đến hứa hẹn cho những ai yêu thích, tò mò về lịch sử và văn hóa người Việt giúp mở mang kiến thức, tăng cường nhận thức và tăng khả năng phát triển tư duy của con người thông qua những trải nghiệm du lịch.

Theo số liệu thống kê từ Tổng cục Du lịch, năm 2022, sau khi gỡ bỏ mọi hạn chế đi lại, lượng khách nội địa cả năm đã đạt 101,3 triệu lượt, tăng 168,3% so với mục tiêu 60 triệu và vượt con số của 2019. Riêng ba tháng hè, lượng khách đạt hơn 35 triệu. Doanh thu đạt 495.000 tỷ đồng, vượt 23% kế hoạch vì thế mà nhiều chuyên gia khẳng định, đây là năm phục hồi của du lịch trong nước (Nam, 2022). Điều này tạo ra một nhu cầu ngày càng tăng về khách du lịch nói chung và khác du lịch đến để trải nghiệm văn hóa lịch sử nói riêng. Bằng những nghiên cứu sơ bộ, nhóm nghiên cứu nhận thấy ý nghĩa trong việc tìm hiểu những điểm đến mang tính lịch sử ở nước ta, nơi có thể giúp cho ta hiểu rõ hơn về chính mình và thế giới xung quanh mình, từ đó giúp ta trở thành một người có trí tuệ và kiến thức văn hóa phong phú hơn.

**2. Tính cấp thiết của đề tài**

Sinh thời, Chủ tịch Hồ Chí Minh đã dạy: “Dân ta phải biết sử ta/Cho tường gốc tích nước nhà Việt Nam”, câu nói đó của Bác thấy được ý nghĩa lịch sử trong vai trò chủ đạo trong giáo dục lòng yêu nước, tinh thần tự tôn dân tộc của nước ta. Tuy nhiên, thực trạng dạy và học môn lịch sử của nước ta hiện nay luôn trong tình trạng đáng báo động với nhiều lý do và trong đó có nguyên nhân là việc học lý thuyết nhiều, và thiếu kết nối với cuộc sống. Vì thế mà việc học lịch sử thực tế qua các địa điểm, sự kiện và nhân vật sẽ giúp giới trẻ cảm thấy hứng thú hơn vì thế mà du lịch lịch sử góp phần mang ý nghĩa quan trọng vì nó giúp chúng ta hiểu rõ hơn về quá khứ của đất nước, văn hóa và con người. Ở đây du khách có thể đến thăm các khu di tích lịch sử, bảo tàng và địa điểm văn hóa để học hỏi, đánh giá và tôn vinh những cống hiến của các bậc tiền bối. Thế nên việc tìm hiểu những mong muốn, những đặc trưng, những khía cạnh và cảm xúc của du khách thông qua bình luận thực tế về các địa điểm có yếu tố lịch sử là rất quan trọng để tìm ra và phát hiện những quan điểm thú vị, hy vọng góp phần phát triển tình hình chung của các địa điểm tham quan văn hóa lịch sử. Vì du lịch lịch sử không chỉ giúp chúng ta học về lịch sử mà điều này còn giúp ta hiểu được cách mà nền kinh tế và văn hóa phát triển, những vấn đề kinh tế, chính trị, văn hóa, xã hội đã được giải quyết trong quá khứ, từ đó học hỏi và áp dụng cho hiện tại và tương lai.

## **3. Tổng quan nghiên cứu**

Các nghiên cứu liên quan nước ngoài:

1. (Michela Fazzolari, Marinella Petrocchi, 2018)

Bài báo phân tích dựa trên 2 tập dữ liệu lấy từ website tripadvisor.com và booking.com là các bình luận đánh giá của người dùng về khách sạn, nhà hàng và điểm đến, thu thập hơn 7 triệu đánh giá. Các phương pháp thực hiện bao gồm: các phương pháp phân tích thống kê để mô tả dữ liệu, kỹ thuật xử lý ngôn ngữ tự nhiên để trích xuất và so sánh những từ xuất hiện với tần suất nhiều và thuật toán khai phá luật kết hợp để điều tra, đánh giá xem các địa điểm nào thường hay xuất hiện cùng nhau. Thông qua các kết quả phân tích nhìn thấy được những đặc điểm, đặc trưng của người dùng từ đó có thể cung cấp thông tin hữu ích cho các nhà cung cấp dịch vụ du lịch để cải thiện chất lượng dịch vụ của họ. Bài báo sử dụng thuật toán khai phá luật kết hợp đã phát hiện ra những điểm đến được người dùng kết hợp với nhau với tần suất cao, là tài liệu hữu ích để đưa ra các khuyến nghị sau này.

Tuy nhiên, bài báo còn có những hạn chế như là: dữ liệu chỉ tập trung ở một Tỉnh nên chưa thấy được sự đặc biệt, đặc trưng thông qua ý kiến người dùng khi đến đây, cùng với đó là chưa đánh giá được thái độ, cảm xúc của người dùng để có thể nhìn sâu hơn sự quan tâm của người dùng.

1. (Viriya Taecharungroj, Boonyanit Mathayomchan, 2019)

Trình bày kết quả phân tích các đánh giá của du khách trên trang TripAdvisor về các điểm tham quan ở Phuket, Thái Lan. Bài báo này sử dụng phương pháp phân tích ngữ liệu và phân tích cảm xúc để hiểu những gì mà khách hàng thực sự nghĩ về các địa điểm tham quan. Các phương pháp nghiên cứu được áp dụng bao gồm: Tiền xử lý dữ liệu, áp dụng thuật toán LDA để tìm ra các từ đặc trưng của từng loại địa điểm tham quan, kết hợp với mô hình phân loại Bayes để phân loại cảm xúc cho các từ đặc trưng dựa vào điểm đánh giá của người dùng.

Kết quả cho thấy rằng những địa điểm tham quan được đánh giá tích cực nhất là các khu vui chơi giải trí, nhà hàng và các địa điểm thể thao ngoài trời, trong khi các địa điểm như đền tưởng niệm, công viên và các địa điểm tham quan liên quan đến văn hóa thường được đánh giá thấp hơn.

Tuy nhiên, việc phân tích điểm số cho các từ trong chủ đề chỉ dựa trên đánh giá người dùng làm sẽ làm cho bài toán không được khách quan, bởi vì người dùng có xu hướng đánh giá cao hơn khiến cho sự xuất hiện của từ ngữ tiêu cực sẽ bị bỏ qua. Cùng với đó là chưa đánh giá được những nhân khẩu học cơ bản của người dùng.

## **4. Mục tiêu nghiên cứu**

Đề tài nghiên cứu ***“ Phân tích dữ liệu để đánh giá sự quan tâm của khách du lịch về các địa điểm tham quan lịch sử ở Việt Nam”*** hướng tới mục tiêu:

Có cái nhìn tổng quan về du lịch lịch sử Việt Nam thông qua dữ liệu và bằng các phương pháp phân tích dữ liệu khác nhau để đưa ra các đặc trưng về các điểm đến văn hóa lịch sử Việt Nam.

Giúp cho người Việt Nam và du khách quốc tế hiểu rõ hơn về lịch sử và văn hóa của ngườiViệt. Điều này giúp cho việc bảo tồn và phát huy giá trị văn hóa của Việt Nam được thực hiện hiệu quả hơn.

Từ dữ liệu văn bản được người dùng bình luận có thể tìm ra các tiềm ẩn, khía cạnh để phát triển, cải thiện hơn chất lượng phục vụ và hiểu rõ hơn về những yêu cầu, mong muốn của khách du lịch.

Điểm đến nào trong Việt Nam có tiềm năng phát triển du lịch lịch sử, văn hóa và danh lam thắng cảnh? Lịch sử và văn hóa Việt Nam đã ảnh hưởng đến các hoạt động du lịch như thế nào? Khách du lịch quốc tế có đánh giá thế nào về các sản phẩm du lịch lịch sử, văn hóa và danh lam thắng cảnh tại Việt Nam?

Các hoạt động du lịch lịch sử, văn hóa và danh lam thắng cảnh ở Việt Nam đang thu hút được những đối tượng du khách nào? Các đối tượng du khách quốc tế và trong nước đang tìm kiếm những trải nghiệm du lịch lịch sử, văn hóa và danh lam thắng cảnh gì tại Việt Nam?

## **5. Đối tượng nghiên cứu**

Nghiên cứu sự quan tâm, nhận thức, đánh giá, và hành vi của khách du lịch đối với các điểm đến lịch sử tại Việt Nam. Cũng như các thông tin khác như: quốc tịch, thời gian du lịch, các hoạt động du lịch, và đánh giá của khách hàng về các điểm đến. Đồng thời, so sánh sự quan tâm của khách du lịch trong nước và khách du lịch quốc tế.

Nghiên cứu sự phổ biến và yêu thích của các điểm đến, cũng như những quan điểm để đánh giá sự đặc trưng về các địa điểm tham quan lịch sử đối với một số tỉnh/ thành có các điểm đến văn hóa lịch sử phổ biến nhất ở Việt Nam.

## **6. Phạm vi nghiên cứu**

Nghiên cứu tập dữ liệu bình luận và đánh giá trên website Tripadvisor, dữ liệu từ năm 14/01/2016 - 15/12/2022 , lĩnh vực về du lịch lịch sử việt Nam.

Phân tích mô tả các thông tin mà khách du lịch đánh giá về các điểm đến lịch sử tại Việt Nam như lịch sử, văn hóa, kiến trúc, danh thắng, địa danh, thánh tích, di sản văn hóa và thiên nhiên. Phân tích những nét đặc trưng về du lịch văn hóa lịch sử ở một vài tỉnh phổ biến nhất Việt Nam thông qua bình luận và phân tích quan điểm theo khía cạnh ở tỉnh có lượt bình luận cao nhất.

## **7. Phương pháp nghiên cứu**

Phương pháp nghiên cứu tài liệu: Nghiên cứu các tài liệu liên quan đến phân tích dữ liệu bình luận, đánh giá trực tuyến của khách hàng về du lịch, nghiên cứu các phương pháp xử lý ngôn ngữ tự nhiên để đưa ra cảm xúc và quan điểm của khách du lịch.

Phương pháp thu thập thông tin: dữ liệu được thu thập là những danh lam & thắng cảnh trích xuất từ website (TripAdvisor), thuộc các thể loại nằm trong từ khóa liên quan đến lịch sử: “Đài tưởng niệm”, “Địa điểm lịch sử”, ”Di tích cổ”, ”Nghĩa trang”,.....

Phương pháp phân tích: Xử lý thông tin định lượng bao gồm phân mô tả và khai phá dữ liệu,phân tích số liệu thống kê để đưa ra dự đoán, tính toán các chỉ số, điều chỉnh và phân tích các kết quả được trực quan hóa để tạo ra những hình ảnh, biểu đồ, đồ thị, từ đó mà việc truyền đạt thông điệp sẽ dễ dàng.

Phương pháp mô hình: Xử lý ngôn ngữ tự nhiên tìm ra các chủ đề tiềm ẩn trong tập dữ liệu văn bản, từ đó tìm ra và phân tích khía cạnh của từng chủ đề.

# **B. NỘI DUNG NGHIÊN CỨU**

## **Chương 1: Cơ sở lý luận**:

***Thuật toán LDA*** (Latent Dirichlet Allocation) (David M. Blei, Andrew Y. Ng, Michael I. Jordan, 2003) là một thuật toán học máy trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) được phát triển bởi David Blei, Andrew và Michael Jordan vào năm 2003. LDA được sử dụng để phân tích các tài liệu văn bản và xác định các chủ đề bên trong chúng. LDA giả định rằng mỗi tài liệu trong bộ dữ liệu được tạo ra bởi một hoặc nhiều chủ đề. Mỗi chủ đề là một tập hợp các từ, và mỗi từ có một xác suất xuất hiện trong từng chủ đề khác nhau bằng cách phân tích phân phối xác suất của các từ trong mỗi tài liệu, và tạo ra một mô hình chủ đề cho toàn bộ dữ liệu.

***Word Cloud*** (Tag Cloud, n.d.)hay còn được gọi là Tag Cloud là một hình ảnh hoặc đồ thị biểu diễn các từ và cụm từ được sắp xếp ngẫu nhiên trong đó, những từ xuất hiện nhiều lần sẽ có kích thước lớn hơn và những từ xuất hiện ít lần sẽ có kích thước nhỏ hơn. Word Cloud thường được tạo ra bằng cách sử dụng các công cụ xử lý ngôn ngữ tự nhiên (NLP) để trích xuất các từ và cụm từ từ một văn bản hoặc tập dữ liệu văn bản. Sau đó, các từ này được đưa vào một chương trình tạo Word cloud để tạo ra hình ảnh hoặc đồ thị. Word cloud thường được sử dụng để trực quan hóa dữ liệu văn bản và giúp người dùng dễ dàng nhận ra các từ và cụm từ quan trọng trong một tập dữ liệu văn bản.

***Du lịch lịch sử*** (historical tourism) là một hình thức du lịch tập trung vào việc khám phá và khai thác các địa điểm và di tích lịch sử, nhằm hiểu rõ hơn về quá khứ của một vùng đất, một quốc gia hay một dân tộc. Du lịch lịch sử có thể bao gồm các hoạt động như tham quan các di tích lịch sử, ghé thăm các bảo tàng, kiến trúc cổ và thị trấn cổ, tham gia các lễ hội truyền thống hoặc các sự kiện liên quan đến lịch sử. Du lịch lịch sử được xem là một hình thức du lịch mang tính giáo dục, nhằm tăng cường kiến thức và hiểu biết về lịch sử, văn hóa và truyền thống của một địa phương hoặc một quốc gia. Đây cũng là một hình thức du lịch mang tính bảo tồn văn hóa, giúp bảo vệ và duy trì các di sản lịch sử, kiến trúc cổ, nghệ thuật và truyền thống của một dân tộc.

***Tripadvisor*** (Tripadvisor, n.d.)là một trang web đánh giá du lịch trực tuyến được thành lập vào năm 2000. Trang web cho phép người dùng đăng nhập và đánh giá các khách sạn, nhà hàng, điểm tham quan và hoạt động du lịch trên toàn thế giới. Các đánh giá này được chia sẻ công khai và được sắp xếp theo độ uy tín và độ hữu ích của người đánh giá. Ngoài ra, Tripadvisor còn cung cấp các chuyến đi, chương trình khuyến mãi và các gợi ý về các điểm đến phổ biến. Nó cũng có tính năng đặt phòng khách sạn và đặt vé máy bay. Tripadvisor hiện là một trong những trang web du lịch lớn nhất thế giới với hơn 500 triệu đánh và đánh giá cho hơn 7 triệu điểm du lịch trên khắp thế giới.

***Data Scraping*** (Web scraping, n.d.)là quá trình tự động trích xuất dữ liệu từ các trang web công cộng. Kỹ thuật này thường được sử dụng để thu thập thông tin từ các trang web mà không cần phải nhập liệu thủ công. Data scraping thường được thực hiện bằng cách sử dụng các công cụ tự động để tìm kiếm và thu thập dữ liệu từ các trang web. Các công cụ data scraping thường sử dụng các thuật toán để tự động đi qua các trang web và trích xuất dữ liệu theo các quy tắc được xác định trước đó. Các kỹ thuật data scraping phổ biến bao gồm phân tích cú pháp HTML, sử dụng API và sử dụng các công cụ tự động để mô phỏng các hành động của con người trên trình duyệt web. Ngoài ra, Data scraping có thể được sử dụng để thu thập thông tin về sản phẩm, giá cả, đánh giá của khách hàng và các thông tin khác từ các trang web bán hàng điện tử. Nó cũng có thể được sử dụng để thu thập thông tin từ các trang web chính phủ, các trang tin tức và các trang web công cộng khác. Tuy nhiên, việc sử dụng kỹ thuật data scraping cũng có thể vi phạm các quy định về bản quyền và chính sách bảo mật của các trang web.

***Topic modelling*** là một kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được sử dụng để phân tích và đưa ra những chủ đề chính trong tập dữ liệu văn bản và được sử dụng rộng rãi trong nhiều lĩnh vực bao gồm phân tích dữ liệu, tìm kiếm thông tin, xử lý ngôn ngữ tự nhiên và phân tích cảm xúc. Kỹ thuật này giúp tóm tắt nội dung của một tập văn bản lớn bằng cách tìm ra những từ khóa phổ biến nhất trong tập dữ liệu và gom chúng thành các nhóm, được gọi là chủ đề. Khi sử dụng các kỹ thuật này, một tập dữ liệu văn bản có thể được biểu diễn dưới dạng ma trận từ khóa-chủ đề, trong đó mỗi hàng tương ứng với một từ khóa và giá trị trong ma trận thể hiện mức độ liên quan giữa từ khóa và chủ đề.

***Khai phá dữ liệu*** là quá trình tìm kiếm và phân tích thông tin từ các tập dữ liệu lớn để tìm ra các mô hình,quy luật hoặc kiến thức ẩn để có thể giúp dự đoán, giải thích hoặc cải thiện quyết định trong tương lai. Khai phá dữ liệu liên quan đến việc sử dụng các kỹ thuật và công cụ phân tích dữ liệu để khám phá thông tin ẩn, quan hệ và xu hướng có trong dữ liệu.

***Trực quan hóa dữ liệu*** (Data Visualization) là quá trình biến đổi các dữ liệu số hoặc thông tin trừu tượng thành các hình ảnh, biểu đồ, đồ thị, bản đồ hoặc các phương tiện trực quan khác để giúp người dùng dễ dàng hiểu và phân tích các dữ liệu phức tạp. Các kỹ thuật trực quan hóa dữ liệu thường được sử dụng để phát hiện mẫu, tìm kiếm những mối quan hệ và xu hướng cũng như để tạo ra các báo cáo và bản đồ có tính tương tác để giúp người sử dụng tương tác trực tiếp với dữ liệu.

***Python*** là một ngôn ngữ lập trình hướng đối tượng đa năng có cấu trúc dữ liệu cấp cao mạnh mẽ và hệ thông thư viện lớn. Python được phát triển vào những năm 1980 bởi Guido van Rossum tại Hà Lan và được công bố lần đầu vào năm 1991. Python có thể được sử dụng cho nhiều mục đích, bao gồm lập trình máy tính, phát triển web, xử lý dữ liệu, trí tuệ nhân tạo, khoa học dữ liệu và nhiều lĩnh vực khác. Python là ngôn ngữ có cấu trúc rõ ràng, được dùng rộng rãi trong phát triển trí tuệ nhân tạo.

***Tiền xử lý dữ liệu văn bản*** là quá trình xử lý và chuẩn bị dữ liệu văn bản trước khi áp dụng các kỹ thuật khai thác dữ liệu hoặc phân tích văn bản. Quá trình này bao gồm loại bỏ các dữ liệu không cần thiết hoặc không liên quan, tiêu chuẩn hóa văn bản, chuyển đổi văn bản thành định dạng phù hợp cho việc xử lý dữ liệu và các bước tiền xử lý khác như tách từ, chuyển đổi chữ hoa thành chữ thường, phân tích cú pháp văn bản. Tiền xử lý dữ liệu văn bản là một bước quan trọng và cần thiết trong quá trình khai thác và phân tích văn bản.

***Khai thác văn bản (text mining)*** (Text mining- khai phá dữ liệu từ văn bản, n.d.)là quá trình khai thác thông tin từ văn bản tự động bằng cách sử dụng các phương pháp và kỹ thuật của khoa học dữ liệu, máy học và xử lý ngôn ngữ tự nhiên. Mục đích của khai thác văn bản là tìm ra các mẫu, quy luật và thông tin tiềm ẩn trong tập dữ liệu văn bản lớn để dự báo kết quả và đưa ra quyết định. Các công cụ và kỹ thuật được sử dụng trong khai thác văn bản bao gồm xử lý ngôn ngữ tự nhiên, phân tích cú pháp, phân tích tương đồng văn bản, phân tích dữ liệu cấu trúc và các kỹ thuật khai thác dữ liệu khác.

***Tableau*** (Tableau là gì?, 2019)là một phần mềm trực quan hóa dữ liệu (data visualization) và phân tích dữ liệu (data analysis) được phát triển bởi Tableau Software. Nó cho phép người dùng kết nối với nhiều nguồn dữ liệu khác nhau, ví dụ như là các cơ sở dữ liệu quan hệ, các tập tin dữ liệu Excel, các tập tin CSV và nhiều nguồn dữ liệu khác. Sau khi kết nối đến dữ liệu, người dùng có thể tạo ra các biểu đồ, bản đồ và bảng điều khiển trực quan để phân tích và hiển thị dữ liệu. Tableau cung cấp một giao diện trực quan và dễ sử dụng cho người dùng, cho phép họ tạo ra các biểu đồ và bảng điều khiển một cách nhanh chóng và dễ dàng mà không cần có kỹ năng lập trình.

***Google Colab*** (viết tắt của Google Colaboratory) là một phiên bản lưu trữ trên đám mây, là một ứng dụng miễn phí của Google cho phép người dùng chạy code Python trong môi trường Jupyter notebook trực tuyến trên các máy chủ của Google.Với Google Colab, người dùng không cần cài đặt các phần mềm và thư viện trên máy tính của mình, mà có thể sử dụng các nguồn tài nguyên tính toán của Google để thực thi code, bao gồm CPU, GPU và TPU. Google Colab cũng cho phép người dùng chia sẻ các notebook của mình với người khác.

***Scrapy*** (Tổng quan kiến trúc của Scrapy, 2020)là một framework mã nguồn mở cho việc crawl và scrape dữ liệu từ các trang web. Nó được viết bằng Python và cung cấp cho người dùng một cách tiếp cận dễ dàng và mạnh mẽ để truy cập và lấy dữ liệu từ các trang web. Scrapy cung cấp một số tính năng hữu ích như xử lý đa luồng, hỗ trợ cho các giao thức HTTP, đa phương tiện và WebSocket, phân tích cú pháp HTML và XML, cùng với các công cụ phân tích dữ liệu và lưu trữ dữ liệu. Với Scrapy, người dùng có thể tạo các spider (con nhện) để tự động crawl và scrape các trang web, thu thập thông tin và lưu trữ dữ liệu vào các cơ sở dữ liệu hoặc file khác nhau. Scrapy được sử dụng rộng rãi trong các ứng dụng web scraping, đánh giá sản phẩm, dữ liệu phân tích và khai thác dữ liệu.

***Based-aspect sentiment analysis*** (Chencheng Ma, Xuehui Du, Lifeng Cao, 2019) là phương pháp phân tích cảm xúc được sử dụng để xác định những phần khác nhau của một tài liệu (ví dụ: đánh giá sản phẩm hoặc bài đăng trên mạng xã hội) và đưa ra đánh giá cảm xúc riêng cho mỗi phần đó. Điều này cho phép người dùng đánh giá cảm xúc về một sản phẩm hoặc dịch vụ theo từng khía cạnh cụ thể, chứ không chỉ nhìn chung về sản phẩm hoặc dịch vụ đó. Ví dụ, trong một đánh giá sản phẩm, based-aspect sentiment analysis có thể phân tích và đưa ra đánh giá cảm xúc riêng cho các khía cạnh như chất lượng, giá cả, độ bền, thiết kế, tính năng. Điều này giúp người dùng có được cái nhìn tổng thể và chi tiết hơn về sản phẩm hoặc dịch vụ, giúp họ đưa ra quyết định mua hàng hoặc sử dụng dịch vụ tốt hơn. Based-aspect sentiment analysis thường được áp dụng trong lĩnh vực phân tích dữ liệu và xử lý ngôn ngữ tự nhiên, và được sử dụng rộng rãi trong các lĩnh vực như đánh giá sản phẩm, chăm sóc khách hàng, phân tích thị trường.

***Thư viện gensim*** (gensim – Topic Modelling in Python, n.d.)là một thư viện mã nguồn mở được viết bằng Python cho việc xử lý ngôn ngữ tự nhiên và mô hình hóa dữ liệu văn bản. Thư viện Gensim cung cấp các công cụ để tiền xử lý dữ liệu văn bản, xây dựng các mô hình vector đại diện cho văn bản, thực hiện các phép tính trên các mô hình đó và tạo ra các chủ đề từ dữ liệu văn bản.Gensim được sử dụng rộng rãi trong các ứng dụng xử lý ngôn ngữ tự nhiên như phân tích tình cảm, phân tích ý kiến, phân loại văn bản và tóm tắt văn bản. Ngoài ra, Gensim cũng cung cấp các công cụ hỗ trợ cho các công việc liên quan đến khai thác dữ liệu, như xây dựng mô hình phân cụm và phân tích đồ thị. Gensim là một thư viện rất mạnh mẽ và linh hoạt, và được sử dụng rộng rãi trong cộng đồng xử lý ngôn ngữ tự nhiên và khai thác dữ liệu.

***Chỉ số perplexity*** là một chỉ số đánh giá mức độ phức tạp của một mô hình ngôn ngữ dựa trên xác suất. Chỉ số perplexity được sử dụng để đánh giá khả năng dự đoán của mô hình trên một tập dữ liệu mới, tức là tập dữ liệu mà mô hình chưa được huấn luyện. Perplexity được tính bằng cách sử dụng entropy của mô hình trên tập dữ liệu mới. Entropy là một thước đo của mức độ bất ngờ của một sự kiện, và trong mô hình ngôn ngữ, nó đo lường khả năng dự đoán của mô hình đối với một từ tiếp theo trong câu. Vì vậy, perplexity càng thấp thì mô hình càng tốt. Chỉ số perplexity thường được sử dụng để so sánh hiệu suất của các mô hình ngôn ngữ khác nhau trên cùng một tập dữ liệu. Nó cũng được sử dụng để đánh giá hiệu quả của các mô hình ngôn ngữ trong các ứng dụng xử lý ngôn ngữ tự nhiên, như dịch máy và sinh văn bản tự động.

***Chỉ số coherence*** (Řehůřek, 2022)là một phương pháp được sử dụng để đánh giá chất lượng của các chủ đề được tạo ra bởi các mô hình chủ đề trong xử lý ngôn ngữ tự nhiên. Chỉ số này đo lường mức độ liên kết giữa các từ được sử dụng trong một chủ đề và đề cập đến mức độ hiểu được của các chủ đề bởi con người. Cụ thể, coherence đo lường khả năng của một tập hợp các từ để hình thành một chủ đề. Nó đánh giá mức độ "thống nhất" giữa các từ trong chủ đề bằng cách tính toán mức độ đồng xuất hiện của các cặp từ trong cùng một văn bản. Các từ được xếp hạng theo mức độ liên kết của chúng, và chỉ số coherence được tính bằng trung bình các xếp hạng này. Chỉ số coherence được sử dụng để đánh giá chất lượng của các chủ đề được tạo ra bởi các mô hình chủ đề và có thể được sử dụng để so sánh các mô hình khác nhau. Mức độ coherence càng cao, chủ đề càng chất lượng và dễ hiểu hơn. Tuy nhiên, cũng cần lưu ý rằng chỉ số coherence không đánh giá được tính đa dạng của các chủ đề, mà chỉ tập trung vào mức độ liên kết giữa các từ trong chủ đề.

***Matplotlib*** (Hoang, 2019)là một thư viện đồ họa trong ngôn ngữ lập trình Python, được sử dụng để tạo ra các biểu đồ và đồ thị tương tác. Thư viện này cung cấp các công cụ vẽ đồ thị linh hoạt, cho phép người dùng tạo ra các loại biểu đồ khác nhau như biểu đồ đường, biểu đồ thanh, biểu đồ hình tròn, biểu đồ phân tán, và nhiều hơn nữa. Matplotlib là một trong những thư viện đồ họa phổ biến nhất trong Python và được sử dụng rộng rãi trong nhiều lĩnh vực như khoa học dữ liệu, machine learning, xử lý hình ảnh, v.v. Với Matplotlib, người dùng có thể tạo ra các biểu đồ chất lượng cao và dễ dàng tùy chỉnh bằng cách sử dụng các lệnh đơn giản trong Python. Matplotlib cũng cung cấp các giao diện đồ họa cho các framework khác trong Python như Pandas và Seaborn, cho phép người dùng dễ dàng tạo ra các biểu đồ phức tạp từ dữ liệu của mình.Matplotlib là một thư viện mã nguồn mở và có sẵn miễn phí cho mọi người sử dụng.

***Phân tích quan điểm (Sentiment analysis)*** là quá trình phân loại, phân tích và đánh giá tâm trạng hoặc quan điểm của người dùng từ văn bản đầu vào. Thông thường, phân tích quan điểm được sử dụng để đánh giá ý kiến của người dùng đối với sản phẩm, dịch vụ, sự kiện, chủ đề. Phân tích quan điểm sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để phân tích và hiểu ý nghĩa của các từ, cụm từ và câu trong văn bản. Phân tích quan điểm được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm kinh doanh, quảng cáo, chính trị, xã hội. Nó giúp các doanh nghiệp và tổ chức đánh giá hiệu quả của chiến dịch quảng cáo, phản hồi từ khách hàng, và đánh giá xu hướng và quan điểm của công chúng.

***Underthesea*** (Anh, 2018)là một thư viện xử lý ngôn ngữ tự nhiên (NLP) mã nguồn mở cho ngôn ngữ tiếng Việt, được phát triển trên nền tảng Python. Thư viện này cung cấp nhiều tính năng, bao gồm :Tách từ (word segmentation), Tách câu (sentence segmentation), Gán nhãn từ loại (part-of-speech tagging), Trích xuất thông tin thực thể (named entity recognition), Tách cụm từ (chunking), Phân tích cú pháp (parsing), Phân tích tôn giáo (sentiment analysis). Underthesea được phát triển và duy trì bởi nhóm nghiên cứu của Trường Đại học Bách Khoa Hà Nội. Thư viện này đã được sử dụng rộng rãi trong các dự án NLP cho tiếng Việt, và được đánh giá là một trong những thư viện NLP tốt nhất cho tiếng Việt. Underthesea có thể được cài đặt thông qua pip và mã nguồn của nó có thể được tìm thấy trên GitHub.

***Natural Language Toolkit*** (Natural Language Toolkit, 2023) là một thư viện mã nguồn mở được viết bằng ngôn ngữ lập trình Python, cung cấp các công cụ và tài liệu hỗ trợ cho xử lý ngôn ngữ tự nhiên (NLP). NLTK cung cấp các công cụ để tiền xử lý văn bản, phân tích cú pháp, phân tích tên thực thể, trích xuất thông tin, tạo ngữ liệu, và nhiều công cụ khác. NLTK là một trong những công cụ phổ biến nhất trong cộng đồng NLP và được sử dụng rộng rãi trong nghiên cứu và ứng dụng thực tế.

***Phân tích cảm xúc dựa trên từ***(Phân tích cảm xúc là gì?, n.d.)là quá trình xác định và đánh giá các cảm xúc, tâm trạng và ý kiến của một cá nhân hoặc một nhóm người dựa trên các từ ngữ được sử dụng trong văn bản. Thông thường, phân tích cảm xúc dựa trên từ được sử dụng trong các ứng dụng NLP để tự động phân loại các văn bản theo cảm xúc của chúng, ví dụ như xác định một email là tích cực hay tiêu cực, phân tích cảm xúc của các bài đăng trên mạng xã hội hoặc đánh giá sản phẩm. Các kỹ thuật phân tích cảm xúc dựa trên từ bao gồm sử dụng các phương pháp học máy và các công cụ xử lý ngôn ngữ tự nhiên để phân tích và đánh giá các từ trong văn bản để xác định cảm xúc của chúng.

***TextBlob*** (textblob.readthedocs.io, 2020)là một thư viện mã nguồn mở được viết bằng Python, cung cấp các công cụ và tài liệu hỗ trợ cho xử lý ngôn ngữ tự nhiên (NLP). Thư viện này cung cấp các công cụ để tiền xử lý văn bản, phân tích cú pháp, phân tích tình cảm, dịch máy và trích xuất thông tin từ văn bản. Nó sử dụng thư viện nltk để phân tích cú pháp và thư viện Pattern để phân tích tình cảm. TextBlob được xây dựng trên Natural Language Toolkit (NLTK), là một trong những thư viện NLP phổ biến nhất và được sử dụng rộng rãi trong cộng đồng NLP. Với TextBlob, việc xử lý và phân tích văn bản trở nên đơn giản hơn và có thể được thực hiện với một số dòng mã Python.

***Biểu đồ Salience-Valence*** (hay còn gọi là biểu đồ hình bán cầu) được sử dụng trong nghiên cứu cảm xúc để đo lường mức độ quan trọng (salience) và tính chất tích cực hay tiêu cực (valence) của các kích thích. Salience (tạm dịch là "tầm quan trọng") đo lường mức độ thu hút sự chú ý của một kích thích cụ thể.. Biểu đồ Salience-Valence được sử dụng để đo lường cảm xúc của con người đối với các kích thích khác nhau, và được sử dụng rộng rãi trong nghiên cứu nhận thức, tâm lý học và y học.

## **Chương 2: Phân tích dữ liệu:**

### 2.1. Mô tả dữ liệu

Dữ liệu thu thập bằng công cụ Scrapy tại: <https://www.tripadvisor.com/Attractions-g293921-Activities-c47-t2,4,6,7,17,19,26-Vietnam.html>

Dữ liệu từ: 14/01/2016 - 15/12/2022

|  |  |
| --- | --- |
| **Tên biến** | **Giải thích tên biến** |
| Attraction | Tên điểm đến |
| Url | Đường dẫn |
| Rate | Điểm đánh giá trung bình |
| Review number | Số lượng đánh giá |
| Province | Tỉnh/ thành của điểm đến |
| Kind | Loại hình của điểm đến |

***Bảng 1: Data1 gồm 162 địa điểm về du lịch liên quan đến lịch sử***

|  |  |
| --- | --- |
| **Tên biến** | **Giải thích biến** |
| User\_name | Tên người dùng |
| Upvote | Lượt upvote |
| Homeplace | Quê quán người dùng |
| Rate | Đánh giá của người dùng |
| Title | Bình luận tổng quan |
| Trip With | Người đồng hành |
| Content | Bình luận chi tiết |
| Language | Ngôn ngữ |
| Written Time | Thời gian bình luận |
| Triptime | Ngày người dùng tham quan địa điểm đó |

***Bảng 2: Data2 Dữ liệu đánh giá, bình luận người dùng gồm 62 146 dòng***

### 2.2. Trực quan hóa dữ liệu

Sau khi xóa bỏ những dữ liệu bị lặp, dữ liệu được mô tả và trực quan hóa bằng công cụ Tableau:

|  |  |
| --- | --- |
| ***Hình 1. Biểu đồ bong bóng thể hiện tổng số lượt đánh***  **giá ở mỗi tỉnh** | **Hình 2. Biểu đồ map thể hiện phân phối số lượng địa điểm tại từng tỉnh** |

Từ biểu đồ trên có thể thấy, ở miền Nam nổi bật nhất có thành phố HCM, miền Bắc có thủ đô Hà Nội và miền Trung là tỉnh Quảng Nam và Huế với nhiều di sản văn hóa và lịch sử quan trọng, đặc biệt là trong thời kỳ phong kiến, là hai địa điểm lịch sử nổi tiếng ở miền Trung Việt Nam.

Về Hà Nội, nơi có nhiều địa điểm tham quan lịch sử nhất Việt Nam cũng rất dễ hiểu vì Hà Nội là thủ đô của Việt Nam và là một trong những thành phố lâu đời nhất của đất nước, nơi đây ghi dấu nhiều sự kiện quan trọng của đất nước. Vì vậy có nhiều địa điểm tham quan lịch sử ở đây. Và nơi đây cũng có nhiều di sản thế giới được UNESCO công nhận như Văn miếu Quốc tử giám, Thăng Long – Hà Nội Cố Đô, Khu phố cổ Hà Nội.

Về TP HCM, tuy không có nhiều địa điểm tham quan như Hà Nội nhưng lại là nơi chiếm nhiều lượt đánh giá nhất. Hồ Chí Minh là thành phố phát triển nhất Việt Nam cũng như có lịch sử khá phong phú và kiến trúc độc đáo.

|  |  |
| --- | --- |
| **Hình 3. Biểu đồ thể hiện cơ cấu số đánh giá** | **Hình 4. Trung bình đánh giá ở mỗi tỉnh** |

Các địa điểm được trung bình đánh giá là 5 khá ít chỉ chiếm hơn 10% và các địa điểm đa số có đánh giá trung bình là 4 chiếm 37%, tuy vẫn là đánh giá được trang web Tripadvisor cho rằng tích cực, nhưng, sẽ có nhiều nguyên nhân khiến khách du lịch không đánh giá điểm tối đa. Nhìn chung điểm số đánh giá tích cực vẫn là cao hơn gần gấp đôi lượng đánh giá tiêu cực, tuy nhiên đây vẫn là điểm số không quá cao.

Quảng Ninh là nơi có trung bình điểm đánh giá tích cực là cao nhất, và Quảng Ninh cũng là tỉnh có nhiều địa điểm tham quan du lịch đứng thứ 2 ở miền Bắc. Các điểm đến di tích lịch sử đặc biệt ở Tỉnh Quảng Ninh có thể kể đến như là: khu di tích Nhà Trần tại Đông Triều và các di tích lịch sử khác như là: Bạch Đằng, Yên tử, Đền Cửa Ông –Cặp Tiên. Theo (Dung, 2021) vào các dịp tết nguyên đán mỗi năm Quảng Ninh đón khoảng 70-100 vạn du khách, thì 70% trong đó là lượng du khách có mặt tại các di tích, di sản trên địa bàn. Điều này cho thấy, di sản văn hóa đã và đang là thành tố rất quan trọng tạo nên sức hút đặc biệt cho du lịch Quảng Ninh. Qua đây có thể thấy du khách cảm thấy rất hài lòng khi đến tham quan lịch sử ở đây.

Hà Nội có điểm trung bình tương đối cao trong khi 3 tỉnh/thành phổ biến còn lại là thành phố Hồ Chí Minh, Quảng Nam và Huế có trung bình đánh giá khá thấp.

Bình luận thu thập gồm 2 ngôn ngữ:

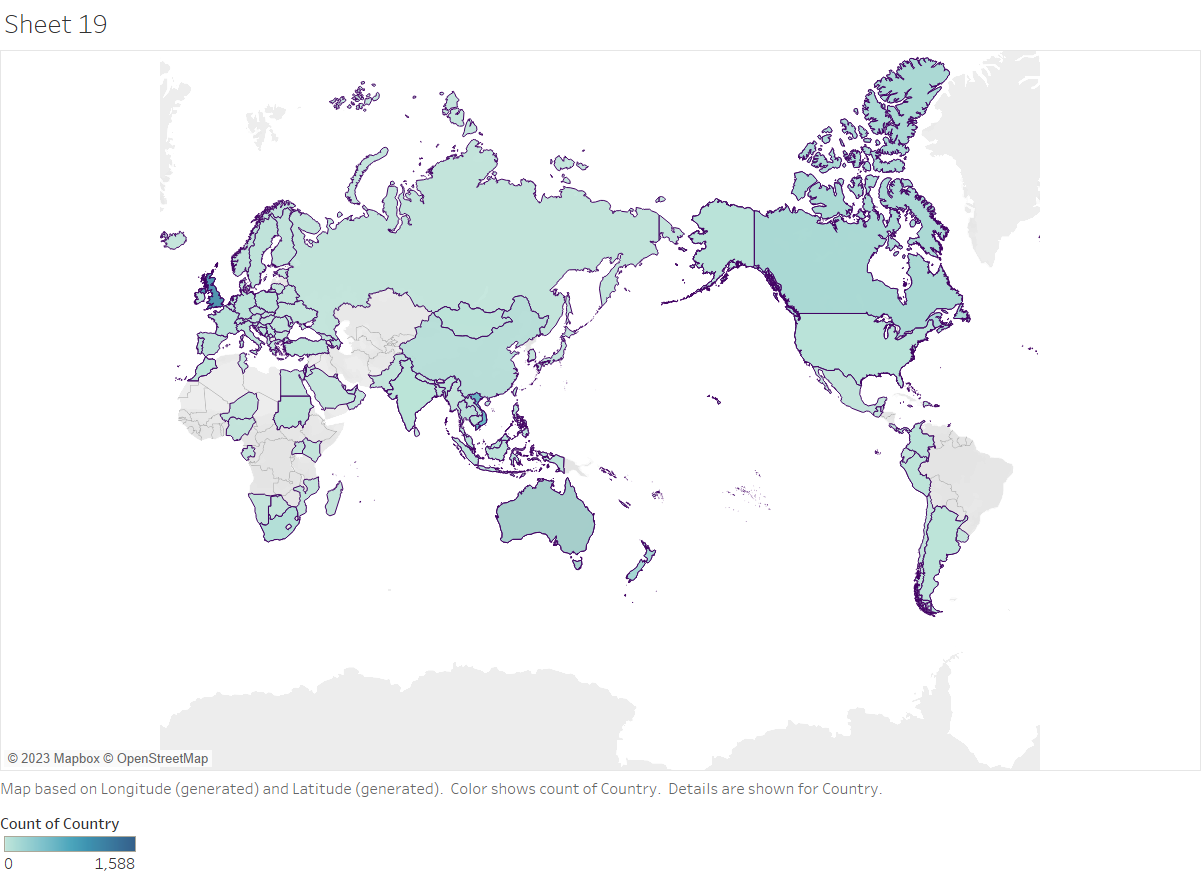
Tiếng Anh: 60,176 bình luận. Dữ liệu về du khách quốc tế.

Tiếng Việt: 1,060 bình luận. Dữ liệu về du khách trong nước.

|  |  |
| --- | --- |
| **Hình 5. Số lượng bài viết qua các năm của từng ngôn ngữ** | **Hình 6 . Loại hình du lịch theo từng ngôn ngữ** |

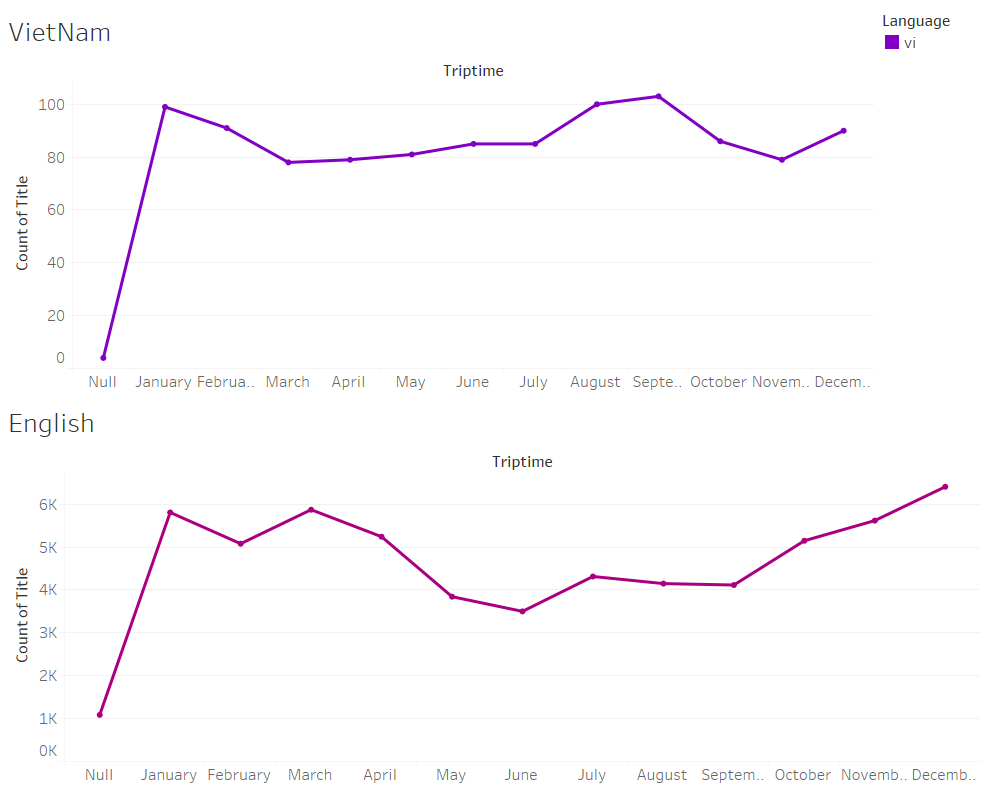
Bình luận người dùng phổ biến từ năm 2015-2019 và giảm đột ngột từ năm 2020 và giảm xuống rất thấp vào năm 2021. Đây cũng là 2 năm mà tình hình dịch bệnh diễn ra rất căng thẳng. Tuy nhiên vào năm 2022 lượng bình lượng có tăng nhưng không nhiều.

Chủ yếu khách quốc tế đến tham quan với hình thức theo cặp đôi chiếm rất cao (hơn 23 nghìn lượt quan sát) và đi cùng với bạn bè. Trong khi du khách Việt Nam sẽ đi cùng với bạn bè và một mình. Với hình thức đi du lịch với doanh nghiệp hoặc với công ty thì không chiếm quá cao.



**Hình 7. Biểu đồ bản đồ thể hiện quốc tịch của người dùng**

Từ biểu đồ có thể thấy lượng khách quốc tế tập trung phần lớn ở phía, Đông và Đông Nam á, Châu Âu và Bắc Mỹ. Ở đây, Úc và Anh chiếm lượng bình luận cao nhất lần lượt là 10,090 và 8,416. Khu vực Đông Á gồm có Singapore, Việt Nam, Malaysia, Ấn Độ, Thái Lan, Trung Quốc. Bắc Mỹ cũng chiếm khá đông gồm Canada và Hoa Kỳ. Biểu đồ thể hiện lượng khách Châu Phi ít hơn hẳn và ở đây chủ yếu là Nam Phi. Vì chỉ lấy lượng người dùng bình luận Tiếng Anh và Tiếng Việt nên dữ liệu khách quốc tế sẽ thiên về những quốc gia sử dụng Tiếng Anh nên phân bố ở trên sẽ không chính xác với tình hình thực tế. Nhưng với lượng khách quốc tế đa dạng từ khắp các quốc gia được thể hiện ở trên thì du lịch lịch sử nước ta cũng khá được quan tâm và phổ biến.



***Hình 8. Biểu đồ đường thể hiện xu hướng thời gian tham quan ở 2 ngôn ngữ***

Đối với khách quốc tế thì thời điểm tháng 12, tháng 1, tháng 3 là khoảng thời gian được du khách ghé thăm nhiều nhất. Đây là những tháng mà nước ta có nhiệt độ dễ chịu mát mẻ, phù hợp với với người nước ngoài.

Đối với du khách Việt Nam, du khách tham quan phổ biến vào tháng 1, tháng 8, và tháng 9 thời điểm hè thu.

Nhìn chung có thể thấy được các địa điểm tham quan lịch sử ở nước ta đón một lượng lớn du khách vào tháng 1.



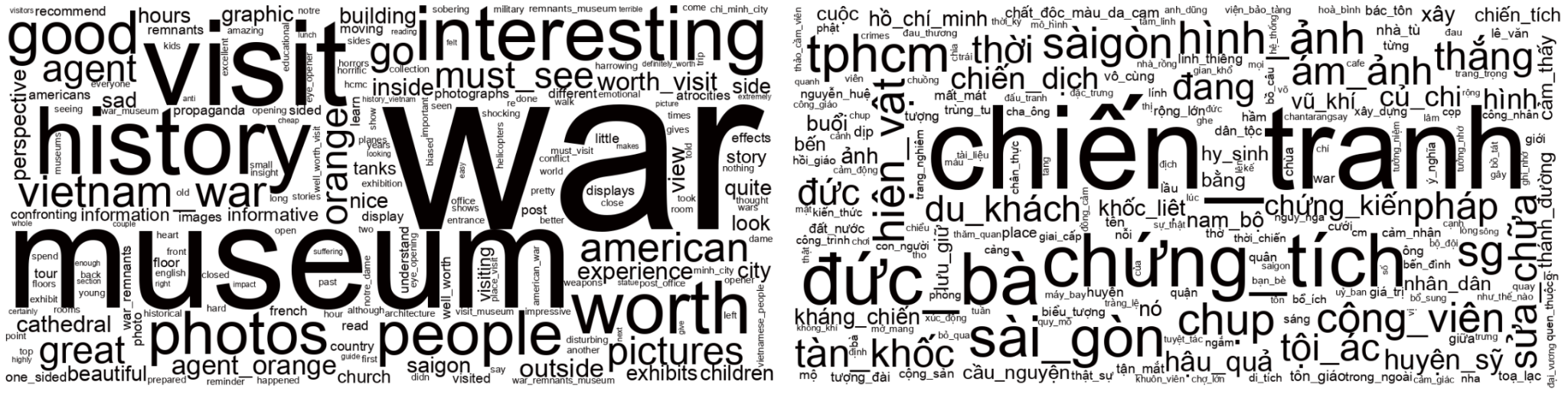
***Hình 9. WordCloud các thể loại của địa đến lịch sử***

Các địa điểm phổ biến bao gồm có nhà thờ, viện bảo tàng, tượng đài, đường phố, ngoài ra còn có những nơi như lăng mộ, di tích cổ, chiến trường,....

### 2.3. Tạo WordCloud cho ba tỉnh phổ biến nhất

WordCloud dựa trên bình luận của du khách về các điểm đến tại 3 tỉnh/thành phổ biến nhất là: Thành phố Hồ Chí Minh, thủ đô Hà Nội, tỉnh Quảng Nam. Tập dữ liệu cũng đồng thời chia ra 2 ngôn ngữ khác nhau để so sánh ý kiến riêng giữa du khách trong và ngoài nước.

**Hồ Chí Minh**

****

***Hình 10. WordCloud bình luận tiếng Việt và tiếng Anh ở thành phố Hồ Chí Minh.***

Tiếng Việt: “Chiến tranh” và “chứng tích” được đề cập phổ biến đi kèm các tính từ như: “tàn khốc”, “ám ảnh”, “tội ác”,”chiến dịch”,”chiến thắng”. Thực tế, trong thời kỳ thuộc địa Pháp và cho đến khi chiến tranh Việt Nam bùng nổ thì Sài Gòn (thành phố Hồ Chí Minh) là nơi đặt trụ sở chính của quân đội Mỹ, trung tâm của chính quyền miền Nam Việt Nam, do đó nhiều địa điểm tham quan ở đây đều liên quan đến các sự kiện trong thời kỳ chiến tranh như như Bảo tàng Chứng tích Chiến tranh, Nhà tù Côn Đảo, Địa đạo Củ Chi,,.... . Bên cạnh đó, nhà thờ Đức Bà - một công trình kiến trúc độc đáo, đẹp mắt với vị trí nằm nằm ở trung tâm của thành phố Hồ Chí Minh, được người Việt đề cập khá nhiều. Ngoài ra còn có 1 số từ khác như là “du khách”,”sửa chữa”, “công viên”,...

Tiếng Anh: từ được đề cập nhiều là “chiến tranh” và “viện bảo tàng”, nhưng những tính từ liên quan khác sẽ là “tốt”, “thú vị”, “đáng giá”, “nên đến”. Du khách đến đây cũng đề cập đến những sự kiện liên quan đến chiến tranh như “người Mỹ”, “chất độc màu da cam”. Ngoài ra còn xuất hiện 1 số từ khác như là “hình ảnh”, “con người”, “khía cạnh”,....

**Hà Nội**

****

***Hình 11. WordCloud bình luận tiếng Việt và tiếng Anh ở thủ đô Hà Nội***

Tiếng Việt: những địa điểm được đề cập phổ biến gồm có “phố phường”, “”hồ gươm”, “hoàn kiếm” với những tính tình liên quan như là “tấp nập”, “nhộn nhịp”,....Ngoài ra ẩm thực đặc trưng ở đây cũng được nhắc đến như : “phở”, “bún”. Không như thành phố Hồ Chí Minh, các điểm đến ở đây gắn liền với văn hóa truyền thống và các kiến trúc cổ.

Tiếng Anh: “đường phố” và “khu phố cổ” là hai địa điểm được nhắc đến nhiều nhất, với các tính từ liên quan như “thú vị”, “tuyệt vời”, “nhộn nhịp”, “tình trạng giao thông”. Ngoài ra từ “đi bộ” cũng nhắc đến khá nhiều, du khách đến đây có vẻ khá chuộng hình thức đi bộ để tham quan. Ngoài ra, cũng đề cập đến những từ như là: “con người”, “ẩm thực”, “hàng hóa”,....

**Quảng Nam**

****

**Hình 12. WordCloud bình luận tiếng Việt và tiếng Anh ở tỉnh Quảng Nam**

Tiếng Việt: “Hội An” và “bánh xèo” được đề cập nhiều cho thấy sự phổ biến của phố cổ “Hội An” và món đặc sản nơi đây như là “bánh xèo”,”nem lụi” . Ngoài ra những từ liên quan đến làng nghề thủ công truyền thống, và những kiến trúc ở cổ ở đây

Tiếng Anh: kiến trúc và công trình cổ được đề cập phổ biến, ngoài ra còn có “viện bảo tàng”, “làng” , các tính từ được miêu tả ở đây gồm có “nhỏ nhắn”, “địa phương”, “thú vị”, “tuyệt vời”, “cổ kính”. Ngoài ra những “tour” du lịch ở đây cũng khá được quan tâm, từ “thuyền” được đề cập nhiều nhắc đến đến hoạt động giải trí như đi thuyền trên sông Thu Bồn ở đây.

## **Chương 3: Phân tích xúc theo khía cạnh**

Dữ liệu: vì dữ liệu hiện tại khá lớn và để chủ đề được tập trung và cụ thể hơn thì nhóm sẽ lọc dữ liệu bình luận có điểm đến thuộc thành phố Hồ Chí Minh (nơi có lượng bình luận cao nhất và lượt đánh giá đa dạng) để quan sát, và phân tích.

Phân tích cảm xúc dựa trên từ vựng gồm các bước sau:

**Bước 1. Tiền xử lý dữ liệu:**

* Xóa những dòng bị trống và bị lặp
* Tạo trường mới kết hợp cả 2 cột title và content, đây sẽ là trường dữ liệu mà nhóm sẽ tiến hành xử lý văn bản
* Dịch ngôn ngữ tiếng Việt thành tiếng Anh

**Bước 2. Tiền xử lý dữ liệu văn bản**

* Tạo và cài đặt bộ từ điển cho văn bản
* Áp dụng thư viện gensim với chức năng là rút trích tính tương tự của văn bản với ngữ liệu lớn cũng như hỗ trợ xây dựng mô hình LDA.
* Tạo các bigram và trigram cho quá trình xử lý để từ vựng được chính xác và các từ kết hợp được liên kết với nhau
* Loại bỏ từ dừng (stop word)
* Chuyển đổi chữ hoa thành chữ thường
* Kỹ thuật Stemming để biến đổi 1 từ về dạng chuẩn
* Kỹ thuật Lemmatization để gán nhãn dữ liệu thành danh từ, tính từ, động từ, trạng từ. Ở quá trình xử lý chọn chủ đề (topic modelling) nhóm sẽ chọn những từ thuộc danh từ, động từ. Và quá trình xử lý word cloud nhóm sẽ chọn những từ tính từ, động từ và danh từ
* Chuyển đổi những từ viết tắt, những cụm slang thường gặp ở tiếng Anh (ví dụ: n’t, wanna, gonna, altho,....)

**Bước 3**. **Tạo mô hình chủ đề**

* Xây dựng mô hình LDA:
* Tiến hành phân tích chủ đề được người dùng hay đề cập để khám phá các chủ đề ẩn trong tập dữ liệu văn bản bình luận.
* Sau nhiều lần thử nghiệm và quan sát nhóm nhận thấy số lượng chủ đề là 3 là hợp lý và phù hợp với mục tiêu nghiên cứu của nhóm.
* Sau đây là kết quả thu được sau khi sau khi chạy mô hình, gồm 10 từ quan trọng nhất ở 3 topic kèm theo số lượng phân phối của chúng.

*[(0, '0.048\*"visit" + 0.041\*"see" + 0.033\*"place" + 0.026\*"go" + 0.021\*"history"''+0.021\*"take"+0.019\*"time" + 0.018\*"museum" + 0.017\*"look" + ''0.017\*"get"'),*

*(1, '0.032\*"house" + 0.031\*"building" + 0.022\*"architecture" + 0.017\*"build" + ''0.014\*"work" + 0.014\*"use" + 0.013\*"view" + 0.013\*"village + ''0.013\*"locate" + 0.012\*"display"'),*

*(2, '0.029\*"statue" + 0.014\*"boat" + 0.014\*"sell" + 0.012\*"bit" + 0.011\*"side" + ' '0.011\*"leave" + 0.010\*"seem" + 0.010\*"road" + 0.009\*"tell" + ''0.009\*"monument"')]*

* Dựa từ những từ ở trên nhóm đã chia thành các chủ đề sau:

1. Hoạt động tham quan
2. Kiến trúc
3. Điạ điểm

* Đánh giá hiệu quả của mô hình:

Perplexity\_score : -7.182763392573849

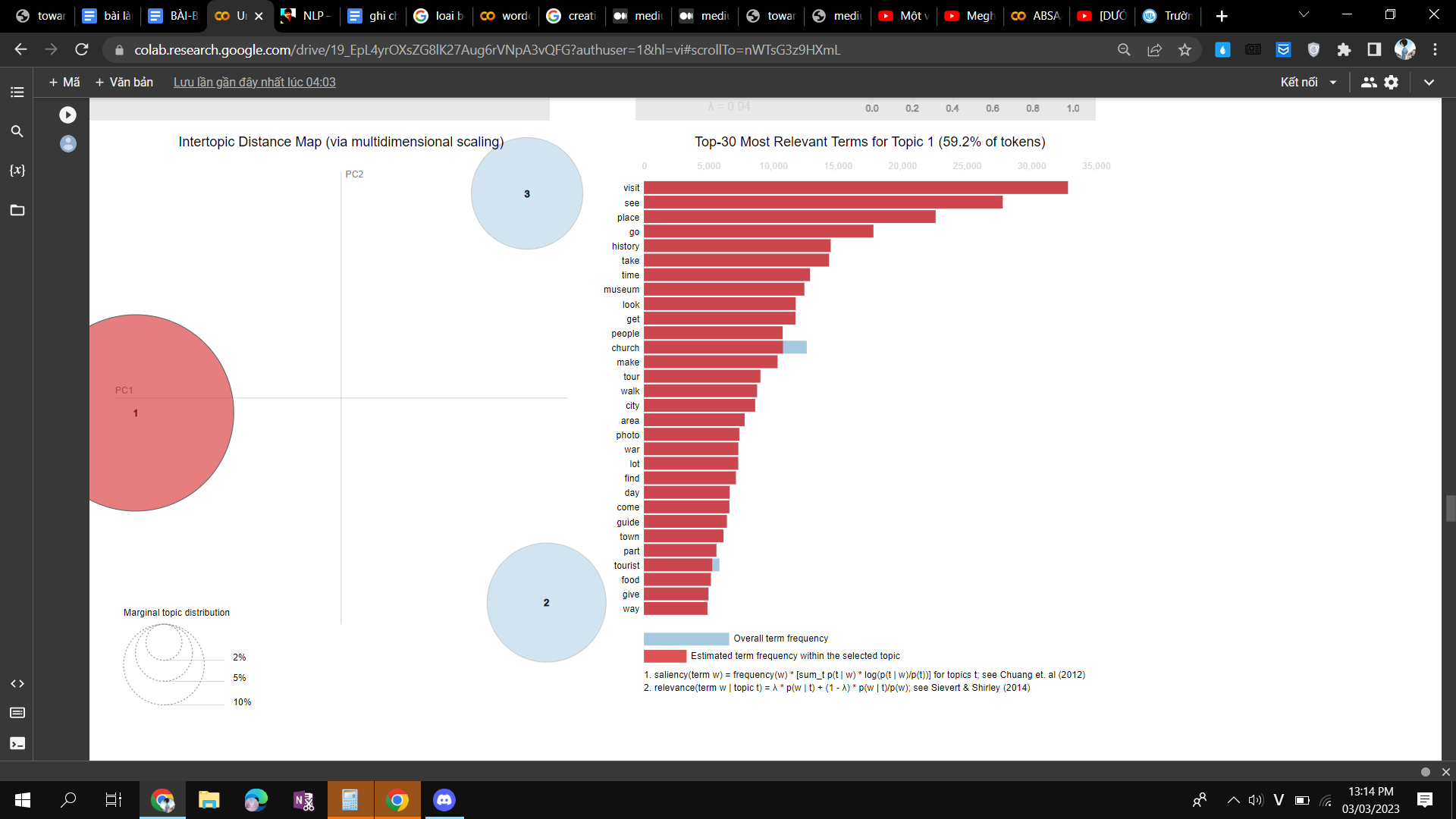
Coherence\_score : 0.5319846603649178

→ Chỉ số perplexity dưới -7 là tương đối thấp , cho thấy mô hình đưa ra dự đoán ổn và có khả năng tổng quát hóa tốt cho văn bản.

→ Chỉ số coherence hơn 0.5 nằm ở mức trung bình, cho thấy các từ trong cùng 1 chủ đề có mối liên kết ổn.

* Trực quan hóa mô hình LDA

Hiển thị kết quả của mô hình LDA:

****

***Hình 13. 30 từ phổ biến nhất ở chủ đề 1***

|  |  |
| --- | --- |
| **Hình 14. 30 từ phổ biến nhất ở chủ đề 2** | **Hình 15. 30 từ phổ biến nhất ở chủ đề 3** |

* Phân tích cảm xúc theo từng chủ đề: Để tiến hành phân tích cảm xúc theo từng chủ đề nhóm tiến hành các bước sau:
* Xác định các từ cẩn quan sát trong tập dữ liệu văn bản theo các chủ đề. Để phân tích được liên quan và phù hợp với đề tài nghiên cứu, nhóm sẽ chọn ra 3 chủ đề với mỗi chủ đề gồm 5 từ bao gồm: chủ đề 1 (visit, place, history, time, museum), chủ đề 2 (house, building, architecture, view, village), chủ đề 3 (statue, boat, sell, road, monument)
* Tách cả văn bản thành những những chuỗi nhỏ theo từng dấu chấm câu hoặc những từ như: “tuy nhiên”, “nhưng”, “mặc dù”.
* Xác định chỉ số cảm xúc ở các từ ở trên bằng thư viện Textblob, và các chỉ số cảm xúc sẽ nằm trong khoảng [-1,+1] từ tiêu cực đến tích cực.
* Trích xuất những mô tả liên quan đến từ đó, với tính năng phân loại chuỗi của spaCy, tự động phân tích cấu trúc ngữ nghĩa của một câu và trích xuất những trạng từ, và tính từ liên quan đến những từ đó.
* Trực quan hóa kết quả bằng biểu đồ Salience-Valience, với Salience đo lường mức độ thu hút và sự chú ý, và Valence đo lường tính chất tích cực hoặc tiêu cực ở một từ. Ở đây sẽ gồm có trục tọa độ 2 chiều. Từ có salience cao sẽ nằm ở vị trí cao trong biểu đồ và ngược lại, trong khi từ có valence tích cực sẽ nằm về phía bên phải, trong khi các từ có valence thấp sẽ nằm về phía bên trái



**Hình 16. Phân phối điểm số cảm xúc ở mỗi từ**

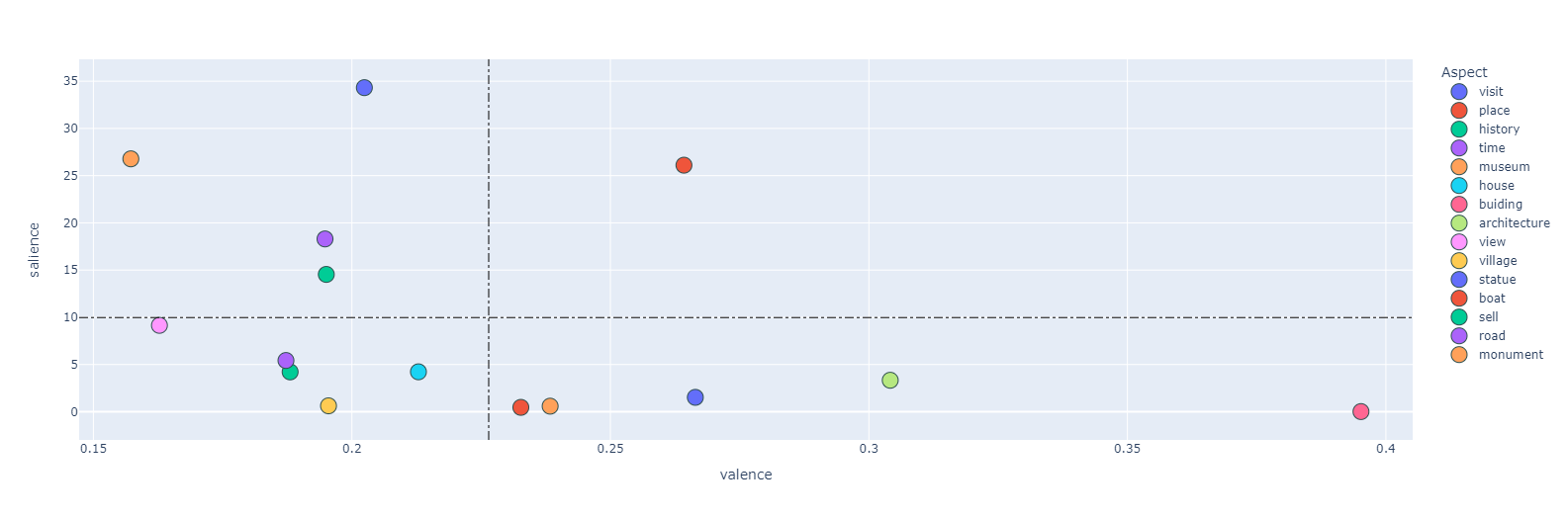
|  |  |  |
| --- | --- | --- |
| **Hình 17. Word Cloud mô tả chủ đề 1** | **Hình 18. Word Cloud mô tả chủ đề 2** | **Hình 19. Word cloud mô tả chủ đề 3** |

Nhìn chung, những chủ đề có tính từ hoặc trạng từ miêu tả khá giống nhau, đều có “thú vị”, “đáng giá” và “tốt”

Chủ đề hoạt động được miêu tả là “cảm xúc”, “thông tin bổ ích” , “buồn”, “xinh đẹp”, “cẩn thận”

Chủ đề về kiến trúc được miêu tả là “sinh động”, “thuộc địa”, “truyền thống”, “cổ kính”, “Pháp”.

Chủ đề về địa điểm có các tính từ như là: “ồn ào”, “thân thiện”, “lộn xộn”, “giá rẻ”, “chật hẹp”, “đầy đủ” và “bất ngờ”



**Hình 20. Biểu đồ Valence - Salience**

Qua vị trí xuất hiện của các từ trên biểu đồ, nhóm có nhận xét về điểm số cảm xúc như sau:

Chủ đề 1: Hoạt động tham quan: 5 từ thuộc chủ đề 1 đều có tầm quan trọng cao nhưng chỉ số cảm xúc không cao lắm, từ “không gian/vị trí” có số điểm cao nhất, từ “viện bảo tàng” có chỉ số tích cực rất thấp nguyên nhân có thể là do những hình ảnh cảm động mà nó mang lại và 3 từ còn lại là “lịch sử”, “thời gian”, “thăm viếng” có chỉ số tích cực gần bằng nhau, không cao không thấp

Chủ đề 2: Kiến trúc: 5 từ thuộc chủ đề 2 có tầm quan trọng tương đối thấp, 2 từ “kiến trúc” và “tòa nhà” có chỉ số cảm xúc tích cực cao nhất trong số các từ. Từ “cảnh quang” được đánh giá là thấp nhất, 2 từ còn lại là “làng” và “nhà” có điểm gần bằng nhau, điểm tích cực không cao không thấp.

Chủ đề 3: Điểm đến: “Tượng” là từ có điểm số tích cực cao nhất trong chủ đề, và 2 từ “bán”, “con đường” có điểm tích cực thấp nhất và gần như bằng nhau, 2 từ còn lại là: “thuyền” và “tưởng niệm” có điểm số tích cực không cao không thấp.

# **C. KẾT LUẬN VÀ KIẾN NGHỊ**

## **1. Kết luận**

Từ những kết quả đánh giá trên thấy được mỗi điểm đến này đều mang lại cho du khách những trải nghiệm khác nhau về văn hóa, lịch sử và đời sống địa phương. Kết quả trực quan đã đưa ra những khám phá thú vị về khách du lịch trong và ngoài nước khi đến với Việt Nam. Bằng cách sử dụng kỹ thuật word cloud đã nhận xét được sự khác biệt, những nét đặc trưng cơ bản về các tỉnh/thành phổ biến tại Việt Nam. Cuối cùng bài nghiên cứu đã đưa ra những khía cạnh cảm xúc khác nhau của khách du lịch về các chủ đề khác nhau tại điểm đến tại thành phố Hồ Chí Minh.

Tuy nhiên, bài báo vẫn chưa bao quát hết về các điểm đến thuộc những tỉnh/thành phổ biến khác tại Việt Nam. Bài toán chưa đánh giá để tìm ra mô hình phù hợp nhất để phân tích cảm xúc và chưa có sự huấn luyện để bài phân tích cảm xúc được chính xác hơn.

## **2. Đề nghị**

Hướng phát triển: Bài nghiên cứu dự kiến áp dụng thêm các kỹ thuật, mô hình khác nhau như: Mô hình Word Embedding áp dụng để biểu diễn các từ trong văn bản bằng một vector số thực, giúp cho việc phân tích cảm xúc trở nên hiệu quả hơn bằng cách tăng cường khả năng hiểu được ngữ nghĩa của từ, hoăc là các thuật toán học máy, học sâu như Convolutional Neural Network (CNN), Recurrent Neural Network (RNN),…. Bằng cách đánh giá, so sánh hiệu suất kết quả mô hình với nhau để có thể áp dụng vào phân tích để bài toán phân tích cảm xúc được chính xác hơn, cho ra nhiều khía cạnh hơn, ở nhiều điểm đến tại các tỉnh/thành khác nhau ở Việt Nam.

# **D. TÀI LIỆU THAM KHẢO**

Anh, V. (2018). *Underthesea documentation*. Retrieved from underthesea.readthedocs.io: https://underthesea.readthedocs.io/en/latest/

Chencheng Ma, Xuehui Du, Lifeng Cao. (2019). Analysis of Multi-Types of Flow Features Based on Hybrid Neural Network for Improving Network Anomaly Detection. *IEEE Access*, 148363-148380.

David M. Blei, Andrew Y. Ng, Michael I. Jordan. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*.

Diem-Trinh Thi Le, Douglas G. Pearce. (2011). Segmenting Visitors to Battlefield Sites: International Visitors to The Former Demilitarized Zone in Vietnam. *Journal of Travel & Tourism Marketing*.

Dung, N. (2021, 11). *Quảng Ninh: Di sản văn hóa - tài nguyên quý để phát triển du lịch*. Retrieved from vietnamtourism.gov.vn: https://vietnamtourism.gov.vn/post/38601

*gensim – Topic Modelling in Python*. (n.d.). Retrieved from github.com: https://github.com/RaRe-Technologies/gensim

Hoang, N. V. (2019, 7 21). *Giới thiệu về Matplotlib* . Retrieved from viblo.asia: https://viblo.asia/p/gioi-thieu-ve-matplotlib-mot-thu-vien-rat-huu-ich-cua-python-dung-de-ve-do-thi-yMnKMN6gZ7P

Mai Ngoc Khuong, Huynh Thi Thu Ha. (2014). The Influences of Push and Pull Factors on the International Leisure Tourists’ Return Intention to Ho Chi Minh City, Vietnam - A Mediation Analysis of Destination Satisfaction. *International Journal of Trade, Economics and Finance*.

Michela Fazzolari, Marinella Petrocchi. (2018). A study on online travel reviews through intelligent data analysis. *Information Technology & Tourism*.

Nam, N. (2022, 12 24). *2022 - năm bùng nổ của du lịch nội địa*. Retrieved from vnexpress.net: https://vnexpress.net/2022-nam-bung-no-cua-du-lich-noi-dia-4551918.html

*Natural Language Toolkit*. (2023). Retrieved from nltk.org: https://www.nltk.org/

*Phân tích cảm xúc là gì?* (n.d.). Retrieved from aws.amazon.com: https://aws.amazon.com/vi/what-is/sentiment-analysis/#:~:text=Ph%C3%A2n%20t%C3%ADch%20c%E1%BA%A3m%20x%C3%BAc%20l%C3%A0%20qu%C3%A1%20tr%C3%ACnh%20ph%C3%A2n%20t%C3%ADch%20v%C4%83n,x%C3%A3%20h%E1%BB%99i%20v%C3%A0%20%C4%91%C3%A1nh%20gi%C3%A1

Pornpisanu Promsivapallop, Prathana Kannaovakun. (2017). A comparative assessment of destination image, travel risk perceptions and travel intention by young travellers across three ASEAN countries: a study of German students. *Asia Pacific Journal of Tourism Research*.

Řehůřek, R. (2022). *Evaluating topic models*. Retrieved from radimrehurek.com: https://radimrehurek.com/gensim/wiki.html#evaluating-topic-models

*Tableau là gì?* (2019, 12). Retrieved from bacs.vn: https://www.bacs.vn/vi/blog/cong-cu-ho-tro/tableau-la-gi-nhung-dieu-can-biet-ve-tableau-data-visualization-6145.html

*Tag Cloud*. (n.d.). Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Tag\_cloud

*Text mining- khai phá dữ liệu từ văn bản*. (n.d.). Retrieved from websitehcm.com: https://websitehcm.com/text-mining-khai-pha-du-lieu-tu-van-ban/

textblob.readthedocs.io. (2020). *TextBlob: Simplified Text Processing.* từ https://textblob.readthedocs.io/en/dev/#.

Thu Thi Trinh, Chris Ryan. (2013). Museums, exhibits and visitor satisfaction: a study of the Cham Museum, Danang, Vietnam. *Journal of Tourism and Cultural Change*.

*Tổng quan kiến trúc của Scrapy*. (2020, 06 05). Retrieved from nordiccoder.com: https://nordiccoder.com/blog/tong-quan-kien-truc-cua-scrapy/#:~:text=Scrapy%20l%C3%A0%20m%E1%BB%99t%20Python%20Framework,v%C3%A0%20extract%20data%20t%E1%BB%AB%20ch%C3%BAng

*Tripadvisor*. (n.d.). Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Tripadvisor

Viriya Taecharungroj, Boonyanit Mathayomchan. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*.

*Web scraping*. (n.d.). Retrieved from wikipedia.org: https://en.wikipedia.org/wiki/Web\_scraping