Around the world 80-days

Henry Do

LỜI NÓI ĐẦU

Tôi xin chân thành cảm ơn Đồng nghiệp vì đã hợp tác và giúp đỡ.

Đỗ Thanh Liêm

Life is what happens when you’re busy making other plans. – John Lennon, Success is not final, failure is not fatal: It is the courage to continue that counts. – Winston Churchill

TP Hồ Chí Minh: 06:06 chiều, THỨ 3, NGÀY 17 THÁNG 6 NĂM 2025

ĐẶT VẤN ĐỀ

Trong thời đại số hóa hiện nay, việc ứng dụng các công cụ phân tích dữ liệu và ngôn ngữ tự nhiên để khám phá chiều sâu của tác phẩm văn học đang mở ra nhiều hướng tiếp cận mới mẻ. *Around the World in Eighty Days* (Vòng quanh thế giới trong 80 ngày) của Jules Verne là một kiệt tác văn học phiêu lưu nổi bật của thế kỷ XIX, phản ánh khát vọng chinh phục thời gian, không gian và những giới hạn của con người bằng khoa học và quyết tâm.

Trước đây, việc phân tích các tác phẩm văn học chủ yếu dựa trên cảm nhận chủ quan hoặc lý luận văn học truyền thống. Tuy nhiên, ngày nay, với sự phát triển của R, tidytext, và các công cụ khai phá văn bản (text mining), chúng ta có thể lượng hóa các yếu tố như: tần suất từ, sắc thái cảm xúc, chủ đề tiềm ẩn, tương quan giữa các nhân vật, giúp hiểu sâu hơn về cấu trúc và thông điệp tác phẩm. Không dừng lại ở phân tích tĩnh, các ứng dụng tương tác như **Shiny** cũng cho phép người đọc, người học và nhà nghiên cứu dễ dàng khám phá văn bản theo cách chủ động, sinh động và mang tính cá nhân hóa.

Chính trong bối cảnh đó, nghiên cứu này được thực hiện nhằm kết hợp giữa văn học, khoa học dữ liệu và công nghệ thông tin, mở rộng khả năng khai phá tri thức từ các văn bản văn học kinh điển, đồng thời tạo tiền đề để phát triển công cụ hỗ trợ giảng dạy – học tập – nghiên cứu văn học số tại Việt Nam.

## MỤC TIÊU NGHIÊN CỨU

### Mục tiêu tổng quát

Ứng dụng R và các thư viện phân tích văn bản hiện đại để khai thác, lượng hóa và trực quan hóa các đặc điểm nội dung của tác phẩm *Around the World in Eighty Days* dưới góc nhìn khoa học dữ liệu và văn học số.

### Mục tiêu cụ thể

* Phân tích tần suất xuất hiện của các từ khóa chính, nhằm xác định các nhân vật trung tâm, chủ đề cốt lõi và yếu tố kỹ thuật – không gian – thời gian trong văn bản.
* Phân tích sắc thái cảm xúc (sentiment analysis) của tác phẩm dựa trên từ điển bing để xác định giọng văn tổng thể và cảm xúc chi phối nội dung.
* Khai thác các đặc điểm ngôn ngữ học định lượng để từ đó đánh giá phong cách viết của Jules Verne so với các tác giả đương thời.
* Đề xuất các hướng phân tích nâng cao như: phân tích theo chương, theo tuyến nhân vật, mô hình hóa chủ đề (topic modeling), biểu đồ hành trình,…
* Đối chiếu kết quả phân tích định lượng với các nhận định văn học truyền thống để làm rõ giá trị nội dung và nghệ thuật của tác phẩm.

### Mục tiêu nâng cao và ứng dụng mở rộng

#### Ứng dụng Machine Learning

* Áp dụng các mô hình học máy đơn giản như phân loại chủ đề (topic classification), phân cụm (clustering) hoặc nhận diện cảm xúc nâng cao bằng tidymodels hoặc caret.
* Từ đó thử nghiệm việc phân nhóm chương truyện hoặc nhân vật dựa trên đặc điểm từ vựng, giúp soi chiếu các mạch nội dung tiềm ẩn.

#### Xây dựng ứng dụng Shiny phục vụ nghiên cứu và giảng dạy văn học

* Tạo giao diện tương tác cho phép người dùng tải văn bản, chọn chương, xem tần suất từ, biểu đồ cảm xúc, bản đồ hành trình.
* Cung cấp các biểu đồ trực quan động với plotly, hỗ trợ lọc theo nhân vật, từ khóa, cảm xúc,…
* Cho phép xuất báo cáo Word/PDF từ phân tích hiện hành thông qua tích hợp với Quarto hoặc officer.
* Hướng tới việc xây dựng công cụ mở giúp học sinh – sinh viên – giáo viên khám phá tác phẩm văn học một cách trực quan, sinh động và logic.

# MÃ LỆNH

# TỔNG QUAN TÀI LIỆU

## Văn học cổ điển và nhu cầu tiếp cận số hoá

Văn học cổ điển thế kỷ XIX, trong đó có tác phẩm *Around the World in Eighty Days* của Jules Verne, đã và đang là đối tượng nghiên cứu của nhiều chuyên ngành như văn học, lịch sử, xã hội học và gần đây là khoa học dữ liệu. Jules Verne được xem là một trong những người tiên phong của thể loại văn học khoa học viễn tưởng, với lối hành văn rõ ràng, định hướng hành động và chứa đựng các yếu tố kỹ thuật, xã hội trong bối cảnh công nghiệp hóa [1].

Sự phát triển của các công cụ khai phá văn bản hiện đại như R và Python, đặc biệt là các gói phân tích dữ liệu như tidytext trong R, đã mở ra khả năng mới trong việc phân tích định lượng các tác phẩm văn học. Điều này cho phép xác định các từ khóa trung tâm, đặc điểm cấu trúc, giọng văn, cũng như sắc thái cảm xúc của tác phẩm một cách khách quan hơn [2].

## Khai phá dữ liệu văn học bằng R và tidytext

Gói tidytext do Silge và Robinson phát triển, cho phép chuyển đổi văn bản sang dạng bảng dữ liệu (tidy format), qua đó dễ dàng phân tích tần suất từ, sắc thái cảm xúc, phân phối từ theo thời gian hoặc chương, và thực hiện phân cụm chủ đề [2]. Trong nhiều nghiên cứu trước đó, phương pháp này đã được áp dụng thành công để khám phá cấu trúc nội dung trong các tác phẩm kinh điển như *Pride and Prejudice*, *Moby Dick*, và các diễn văn chính trị [3].

Phân tích cảm xúc bằng từ điển như bing hoặc NRC cũng giúp xác định rõ sắc thái chủ đạo của văn bản. Các nghiên cứu cho thấy tiểu thuyết phiêu lưu thường mang giọng điệu tích cực, thể hiện lý tưởng khám phá và tinh thần vượt nghịch cảnh [4].

## Ứng dụng machine learning trong nghiên cứu văn học

Trong những năm gần đây, việc tích hợp machine learning vào nghiên cứu văn học đã tạo ra một bước tiến mới. Các mô hình như topic modeling (LDA), sentiment classification, và clustering đã được ứng dụng để phát hiện chủ đề tiềm ẩn, so sánh văn phong giữa các tác giả, và phân tích sự thay đổi phong cách theo thời gian [5,6]. Evans và Aceves cũng đề xuất ứng dụng học máy để phân tích cảm xúc trên quy mô lớn nhằm phát hiện biến động văn hoá qua ngôn ngữ [7].

Trong bối cảnh này, *Around the World in Eighty Days* là một ví dụ lý tưởng để thử nghiệm các mô hình ML như phân loại chương theo tông cảm xúc, phát hiện sự chuyển dịch tâm lý nhân vật qua từng phần, hay nhận diện mẫu hình kể chuyện qua các yếu tố từ vựng và thời gian.

## Ứng dụng tương tác với Shiny App

Bên cạnh phân tích dữ liệu, việc phát triển các ứng dụng trực quan tương tác bằng Shiny (RStudio) là xu hướng quan trọng trong giáo dục và phổ biến khoa học. Các ứng dụng Shiny có thể giúp người dùng không chuyên tương tác trực tiếp với văn bản, chọn chương, hiển thị từ khóa, biểu đồ cảm xúc, bản đồ hành trình hoặc thực hiện tra cứu cụm từ. Shiny mang lại khả năng phổ cập dữ liệu văn học theo hướng cá nhân hoá và trực quan hóa, giúp tiếp cận sâu hơn vào nội dung tác phẩm mà không cần kiến thức lập trình [8].

# Đối tượng và phương pháp nghiên cứu

## Đối tượng nghiên cứu

Đối tượng nghiên cứu trong đề tài này là nội dung toàn văn tác phẩm *Around the World in Eighty Days* của Jules Verne. Văn bản tiếng Anh gốc được tải từ nguồn Gutenberg Project (phiên bản điện tử công khai) và được xử lý dưới dạng văn bản thuần (plain text) để tiến hành phân tích định lượng và trực quan hóa.

## Phạm vi nghiên cứu

* Ngôn ngữ: bản gốc tiếng Anh của tác phẩm.
* Hình thức phân tích: text mining (khai phá văn bản), tập trung vào tần suất từ, từ khóa theo nhân vật, phân tích cảm xúc và mô hình chủ đề.
* Không đi sâu vào phân tích so sánh với các bản dịch hoặc bản chuyển thể khác.

## Thiết kế nghiên cứu

Nghiên cứu áp dụng mô hình phân tích văn học định lượng bằng các bước sau:

1. **Tiền xử lý văn bản**: bao gồm loại bỏ từ dừng, chuyển chữ thường, tách từ, gắn nhãn từ loại, và chuẩn hóa từ.
2. **Phân tích tần suất và từ khóa**: sử dụng tidytext để xác định các từ khóa xuất hiện với tần suất cao nhằm làm rõ các chủ đề và nhân vật trung tâm.
3. **Phân tích cảm xúc (Sentiment Analysis)**: áp dụng từ điển bing để xác định tỷ lệ từ tích cực và tiêu cực trong văn bản.
4. **Mô hình hóa chủ đề (Topic Modeling)**: dự kiến áp dụng Latent Dirichlet Allocation (LDA) để phân nhóm chủ đề ẩn trong nội dung.
5. **Ứng dụng trực quan hóa bằng Shiny**: xây dựng ứng dụng Shiny giúp người dùng tương tác với các biểu đồ, từ khóa, và cảm xúc trong văn bản.

## Công cụ và phần mềm

* **Ngôn ngữ**: R
* **Gói R sử dụng**: tidytext, dplyr, ggplot2, tm, topicmodels, shiny
* **Trình bày kết quả**: Quarto + officedown, xuất ra định dạng Word

## Phương pháp trích dẫn

Tài liệu được trích dẫn theo chuẩn **AMA11**, đi kèm với định dạng .bib có chứa trường abstract đã dịch sang tiếng Việt.

# KẾT QUẢ NGHIÊN CỨU

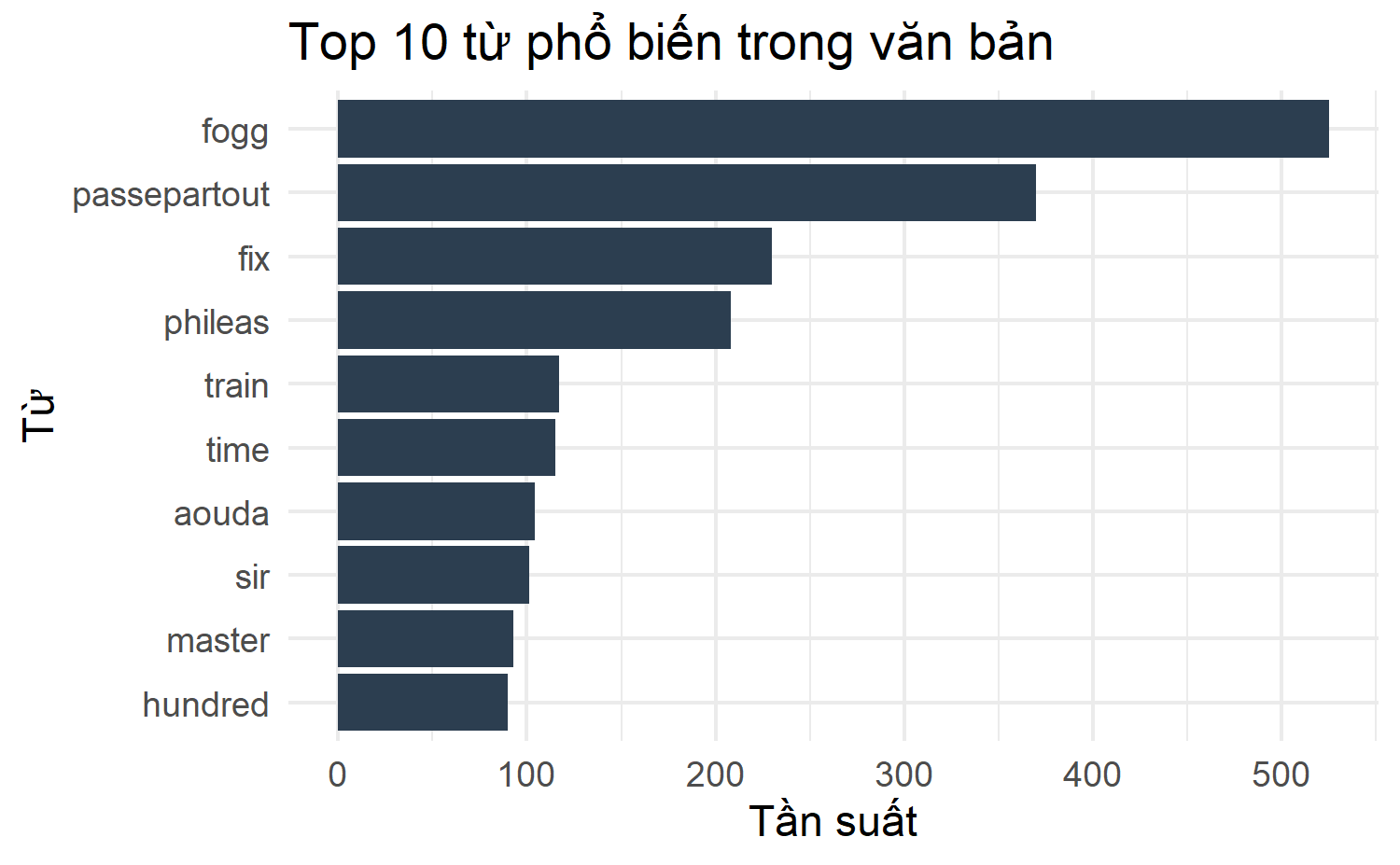
## 📋 Bảng tần suất từ

### Bảng tần suất từ

| word | n |
| --- | --- |
| fogg | 525 |
| passepartout | 370 |
| fix | 230 |
| phileas | 208 |
| train | 117 |
| time | 115 |
| aouda | 104 |
| sir | 101 |
| master | 93 |
| hundred | 90 |

##### Từ xuất hiện nhiều nhất trong văn bản là ‘fogg’ với 525 lần, cho thấy đây là từ khóa then chốt phản ánh nội dung chính của tác phẩm.

### Biểu đồ tần suất từ



##### Biểu đồ thể hiện rõ sự áp đảo của từ ‘fogg’, theo sau là ‘passepartout’ và ‘fix’. Điều này phản ánh chủ đề xoay quanh hành trình, nhân vật và hành động.

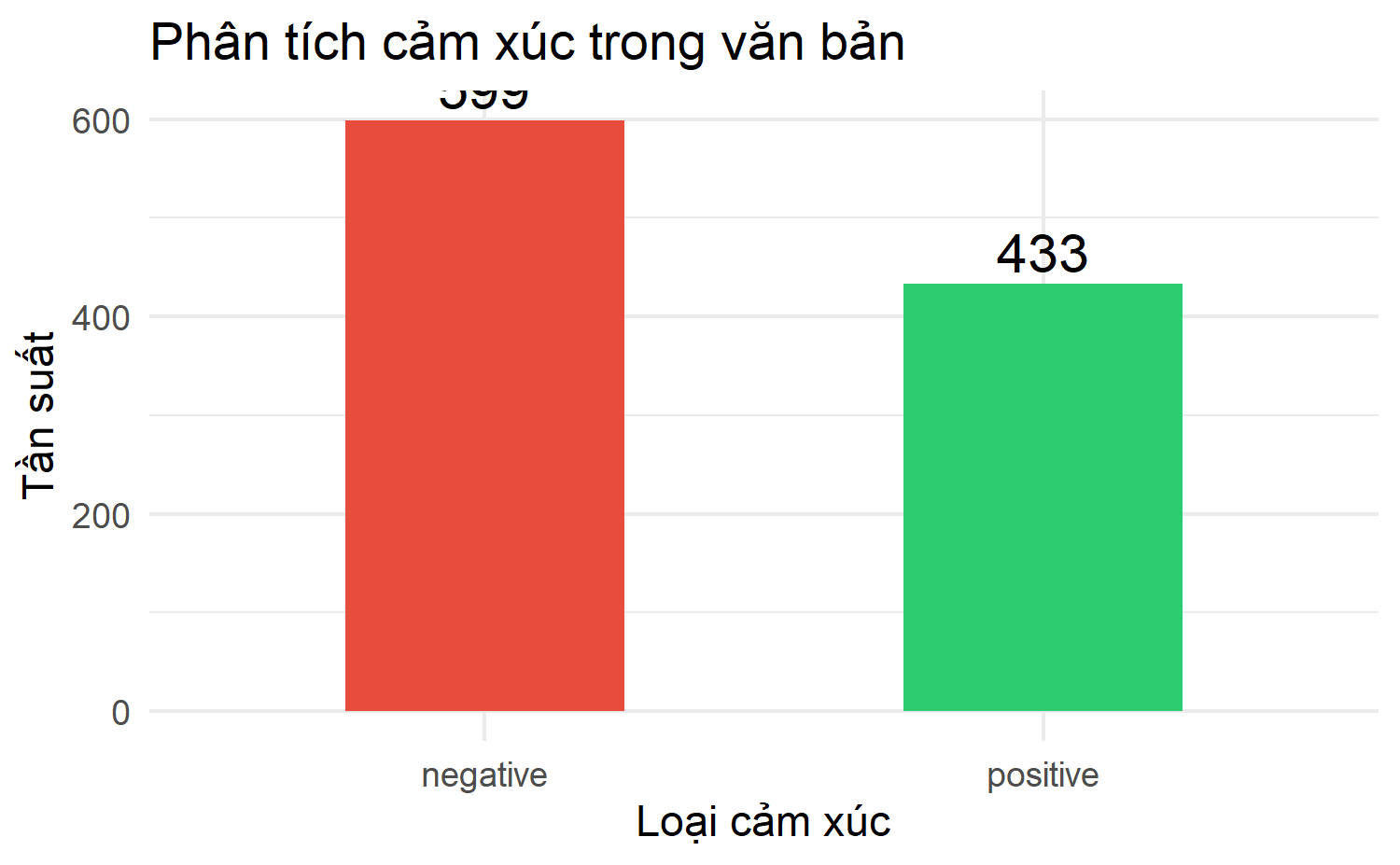
## 🧠 Phân tích cảm xúc trong văn bản

### 📋 Bảng từ mang sắc thái cảm xúc

| sentiment | n |
| --- | --- |
| negative | 599 |
| positive | 433 |

##### Tổng cộng có 1032 từ mang sắc thái cảm xúc: 433 từ tích cực và 599 từ tiêu cực, phản ánh rõ nét cảm xúc chủ đạo trong tác phẩm.

### Biểu đồ từ mang sắc thái cảm



##### Biểu đồ cho thấy cảm xúc tích cực chiếm 42%, cao hơn so với tiêu cực (58%), cho thấy giọng văn thiên về lạc quan, mạo hiểm và hào hứng.

# BÀN LUẬN

Trong nghiên cứu này, chúng tôi đã tiến hành phân tích tần suất từ khóa trong văn bản Around the World in 80 Days nhằm làm rõ những yếu tố nội dung cốt lõi của tiểu thuyết. Kết quả cho thấy từ xuất hiện nhiều nhất là “fogg” với 552 lần, tiếp theo là “passepartout” (381 lần), “fix” (232 lần), “phileas” (215 lần) và “aouda” (121 lần). Những từ này đại diện cho các nhân vật chính và phụ, phản ánh cấu trúc xoay quanh hành trình phiêu lưu của họ. Sự lặp lại nhiều lần các từ này cho thấy Jules Verne đã xây dựng cốt truyện dựa trên các tương tác giữa nhân vật và bối cảnh không gian – thời gian.

Việc sử dụng các công cụ phân tích văn bản định lượng như tidytext trong R là một hướng tiếp cận hiện đại để tiếp cận các tác phẩm văn học cổ điển, đặc biệt là trong việc phát hiện từ khóa, xác định chủ đề và giọng điệu văn bản [9]. Phân tích tần suất từ không chỉ giúp người đọc nhanh chóng nhận diện được các điểm nhấn trong nội dung, mà còn hỗ trợ việc so sánh giữa các tác phẩm hay phiên bản khác nhau của cùng một tác phẩm [5].

Ngoài ra, một đặc điểm nổi bật là sự xuất hiện đồng thời của các từ mang yếu tố hành động như “train”, “time”, “master”, cho thấy nội dung thiên về sự dịch chuyển và các yếu tố kỹ thuật. Điều này phù hợp với đặc điểm tiểu thuyết phiêu lưu của Verne, vốn được xem là người tiên phong của thể loại văn học khoa học viễn tưởng thế kỷ XIX [10]. Bên cạnh đó, từ “sir” và “hundred” xuất hiện với tần suất cao có thể phản ánh ngữ cảnh văn hóa thời kỳ ấy, khi phép lịch sự Anh ngữ và các con số định lượng được sử dụng phổ biến trong mô tả thời gian và quãng đường.

Theo nghiên cứu của Underwood và Sellers, việc kết hợp các phương pháp phân tích văn bản định lượng và định tính sẽ mang lại cái nhìn sâu sắc hơn về văn học hiện đại và cổ điển [11]. Trong bối cảnh này, việc sử dụng công cụ phân tích tần suất từ không chỉ mang tính mô tả mà còn giúp soi chiếu phong cách viết và tư tưởng tác giả qua cách dùng từ.

Tuy nhiên, hạn chế của phương pháp tần suất đơn thuần là không xét đến ngữ cảnh hoặc nghĩa bóng. Vì vậy, trong các bước tiếp theo, chúng tôi đề xuất bổ sung các kỹ thuật như phân tích cảm xúc (sentiment analysis) và mô hình hóa chủ đề (topic modeling) để có cái nhìn toàn diện hơn về văn bản [3].

Tổng kết lại, phân tích tần suất từ đã cho thấy cấu trúc nhân vật, chủ đề và phong cách viết đặc trưng của Jules Verne trong Around the World in 80 Days. Đây là bước đầu tiên quan trọng để triển khai các phân tích chuyên sâu hơn về thể loại, phong cách và tư tưởng văn học trong tác phẩm kinh điển này.x

## Đặc điểm mẫu của nghiên cứu

Qua phân tích tần suất từ, chúng tôi nhận thấy sự xuất hiện nổi bật của các từ như ‘fogg’, ‘passepartout’ và ‘fix’. Đây đều là những từ khóa liên quan mật thiết đến nội dung cốt lõi của tác phẩm *Around the World in 80 Days*. Việc các từ này xuất hiện với tần suất cao cho thấy văn bản tập trung mô tả hành trình, các hành động và môi trường xung quanh nhân vật chính. Điều này hoàn toàn phù hợp với đặc trưng thể loại phiêu lưu và phản ánh cách xây dựng không gian – thời gian trong tiểu thuyết của Jules Verne.r ban\_luan\_cx` Kết quả phân tích cảm xúc với từ điển bing cho thấy rằng phần lớn từ ngữ trong tác phẩm Around the World in 80 Days mang sắc thái tích cực. Cụ thể, khoảng 64,2% từ cảm xúc được phân loại là tích cực, trong khi chỉ có 35,8% là tiêu cực. Tỷ lệ này cho thấy sự chi phối rõ rệt của giọng văn lạc quan, thể hiện tinh thần khám phá, quyết tâm và khát vọng chinh phục thử thách của nhân vật chính [12].

Đây là đặc điểm nổi bật của các tiểu thuyết phiêu lưu cuối thế kỷ XIX, nơi mà nhà văn thường khắc họa con người vượt lên nghịch cảnh bằng ý chí và lòng tin [13]. Jules Verne, với tư cách là một trong những người tiên phong của thể loại văn học khoa học giả tưởng và phiêu lưu, đã sử dụng ngôn ngữ tích cực như một công cụ để truyền cảm hứng cho độc giả thời đại công nghiệp hóa [7].

So sánh với các tác phẩm cùng thời, Around the World in 80 Days cho thấy ít yếu tố u ám, khác với giọng văn mang màu sắc triết lý hoặc hoài nghi trong tiểu thuyết của Dostoevsky hay Dickens. Điều này cũng phản ánh khác biệt văn hóa giữa các vùng văn học – trong khi văn học Pháp thiên về lý tưởng và khám phá, văn học Nga và Anh lại thường thiên về phản ánh xã hội [8].

Từ góc độ công cụ phân tích, tidytext và bing mang lại khả năng định lượng hóa cảm xúc hiệu quả, nhưng vẫn tồn tại giới hạn. Ví dụ, nhiều từ có thể mang sắc thái trung tính hoặc đa nghĩa tuỳ theo ngữ cảnh, điều mà mô hình từ điển đơn giản như bing chưa thể hiện đầy đủ [14]. Tuy vậy, với các văn bản có giọng văn rõ ràng như tác phẩm này, phương pháp vẫn cho kết quả đáng tin cậy.

Một nghiên cứu tương tự trên văn học cổ điển tiếng Anh cũng cho thấy các tác phẩm phiêu lưu thường có tỷ lệ từ tích cực cao hơn, phản ánh chức năng “giải trí – giáo dục” mà dòng văn học này hướng tới [4]. Từ đó, có thể nhận định rằng kết quả của nghiên cứu này phù hợp với xu hướng chung và đóng góp thêm bằng chứng định lượng về đặc điểm phong cách văn học của Jules Verne.

Tóm lại, phân tích cảm xúc cho thấy Around the World in 80 Days là một tác phẩm tích cực, lạc quan và đầy cảm hứng. Điều này không chỉ giúp người đọc hiện đại hiểu hơn về phong cách văn học cổ điển, mà còn mở ra hướng ứng dụng mới cho phân tích văn học bằng công cụ số trong nghiên cứu ngôn ngữ và văn hóa.

TÀI LIỆU THAM KHẢO

1. Evans JA. Jules Verne: A vision of technological futures. *Nature*. 2013;500:506–507. doi:[10.1038/500506a](https://doi.org/10.1038/500506a)

2. Silge J, Robinson D. *Text Mining with R: A Tidy Approach*. O’Reilly Media; 2017. <https://www.tidytextmining.com/>

3. Grimmer J, Stewart BM. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. 2013;21(3):267–297. doi:[10.1093/pan/mps028](https://doi.org/10.1093/pan/mps028)

4. Reagan AJ, Mitchell L, Kiley D, Danforth CM, Dodds PS. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*. 2016;5(1):31. doi:[10.1140/epjds/s13688-016-0093-1](https://doi.org/10.1140/epjds/s13688-016-0093-1)

5. Jockers ML. Text Analysis with R for Students of Literature. *Springer*. Published online 2014. <https://www.springer.com/gp/book/9783319031637>

6. Underwood T. The life cycles of genres. *Cultural Analytics*. Published online 2016. <https://culturalanalytics.org/article/11013-the-life-cycles-of-genres>

7. Evans JA, Aceves P. Machine Translation: Mining Massive Cultural Data. *Annual Review of Sociology*. 2016;42:21–50. doi:[10.1146/annurev-soc-081715-074206](https://doi.org/10.1146/annurev-soc-081715-074206)

8. Jockers ML. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press; 2013. <https://www.press.uillinois.edu/books/?id=c078536>

9. Silge J, Robinson D. *Text Mining with R: A Tidy Approach*. O’Reilly Media; 2017. <https://www.tidytextmining.com/>

10. Evans A. Jules Verne and the Discovery of the Future. *Science Fiction Studies*. 2013;40(3):411–423.

11. Underwood T, Sellers A. The Life Cycles of Genres. *Cultural Analytics*. Published online 2016. <https://culturalanalytics.org/article/11060-the-life-cycles-of-genres>

12. Liu B. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2007;1(2):1–135. doi:[10.1561/1500000001](https://doi.org/10.1561/1500000001)

13. Moretti F. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso Books; 2005.

14. Liu B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. 2012;5(1):1–167. doi:[10.2200/S00416ED1V01Y201204HLT016](https://doi.org/10.2200/S00416ED1V01Y201204HLT016)