

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN TỐT NGHIỆP**  
**NGÀNH KHOA HỌC MÁY TÍNH**

**Đề tài: Xây dựng ứng dụng tóm tắt văn bản tự động bằng ngôn ngữ Python**

Giảng viên hướng dẫn : ThS. Ngô Thị Bích Thúy

Sinh viên thực hiện : Phạm Hữu Hải

Mã sinh viên : 2020607388

Lớp : 2020DHKHMT02\_k15

Hà Nội – 2024

# MỤC LỤC

|   |    |
|---|----|
| DANH MỤC HÌNH ẢNH .....   | 5  |
| LỜI CẢM ƠN .....  | 1  |
| LỜI MỞ ĐẦU.....   | 2  |
| CHƯƠNG 1: KHÁI QUÁT BÀI TOÁN TÓM TẮT VĂN BẢN.....                         | 4  |
| 1.1 Bài toán tóm tắt văn bản tự động .....                                | 4  |
| 1.1.1 Tóm tắt văn bản .....   | 4  |
| 1.1.2. Ứng dụng của tóm tắt văn bản.....                                  | 5  |
| 1.1.3. Phân loại tóm tắt văn bản .....                                    | 6  |
| 1.1.4. Mô hình tóm tắt văn bản.....                                       | 8  |
| 1.1.5. Đánh giá văn bản tóm tắt:.....                                     | 9  |
| 1.1.6. Một số đặc trưng và khó khăn trong tóm tắt văn bản tiếng việt..... | 11 |
| 1.2. Các hướng tiếp cận tóm tắt văn bản .....                             | 12 |
| CHƯƠNG 2: MỘT SỐ NGHIÊN CỨU VỀ TÓM TẮT VĂN BẢN .....                      | 14 |
| 2.1 Tổng quan về tóm tắt nội dung.....                                    | 14 |
| 2.2. Phương pháp tóm tắt văn bản.....                                     | 15 |
| 2.2.1 Mô hình TextRank.....   | 16 |
| 2.2.2 Đồ thị vô hướng.....  | 17 |
| 2.2.3 Đồ thị có trọng số .....  | 18 |
| 2.2.4 Đồ thị hóa văn bản.....   | 19 |
| 2.2.5 Sử dụng TextRank trích xuất từ khóa .....                           | 20 |
| 2.2.6 Giải thuật TextRank .....   | 24 |
| 2.2.7 Các phương pháp tính độ tương đồng .....                            | 25 |
| 2.3 Phương pháp tóm tắt video.....  | 27 |

|  |    |
|--|----|
| 2.3.1. Chuyển đổi video sang văn bản.....                  | 27 |
| 2.3.2. Xử lý văn bản đã chuyển đổi.....                    | 27 |
| 2.4 Tóm tắt văn bản theo hướng chính chọn .....            | 29 |
| 2.4.1. Phương pháp chủ đề đại diện dựa trên tần xuất ..... | 29 |
| 2.4.2. Phương pháp đặc trưng đại diện.....                 | 31 |
| 2.5. Tóm tắt văn bản theo hướng tóm lược .....             | 32 |

### **CHƯƠNG 3: XÂY DỰNG ÚNG DỤNG VÀ KẾT QUẢ THỰC**

|   |           |
|---|-----------|
| <b>NGHIÊM.....</b>                                      | <b>34</b> |
| 3.1. Các thư viện được sử dụng.....                     | 35        |
| 3.1.1. Thư viện Flask .....                             | 35        |
| 3.1.2. Thư viện Numpy.....                              | 37        |
| 3.1.3. Thư viện werkzeug.utils .....                    | 38        |
| 3.1.4. Thư viện tempfile .....                          | 39        |
| 3.1.5. Thư viện underthesea .....                       | 39        |
| 3.1.6. Thư viện Numpy Networkx.....                     | 40        |
| 3.1.7. Thư viện moviepy.editor .....                    | 41        |
| 3.1.8. Thư viện SpeechRecognition .....                 | 41        |
| 3.2. Tổng quan ứng dụng tóm tắt văn bản.....            | 43        |
| 3.3. Cài đặt ứng dụng tóm tắt văn bản .....             | 45        |
| 3.3.1. Mô hình giải quyết bài toán.....                 | 45        |
| 3.3.2. Tiền xử lý văn bản.....                          | 46        |
| 3.3.3. Xây dựng mô hình TextRank .....                  | 47        |
| 3.3.4. Xây dựng ma trận tương đồng.....                 | 48        |
| 3.3.5. Áp dụng thuật toán textrank và tạo tóm tắt ..... | 50        |

|  |    |
|--|----|
| 3.4. Kết quả thực nghiệm .....           | 52 |
| 3.4.1. Môi trường thực nghiệm: .....     | 52 |
| 3.4.2. Quá trình thử nghiệm: .....       | 54 |
| 3.4.3. Dự đoán và hiển thị kết quả:..... | 55 |
| 3.5. Hạn chế và hướng phát triển .....   | 66 |
| 3.5.1. Hạn chế của nghiên cứu: .....     | 66 |
| 3.5.2. Hướng phát triển tiềm năng:.....  | 67 |
| KẾT LUẬN.....                            | 68 |
| TÀI LIỆU THAM KHẢO .....                 | 69 |

## **DANH MỤC HÌNH ẢNH**

|  |    |
|--|----|
| Hình 2.1 Thuật toán PageRank .....                                       | 15 |
| Hình 2.2 Công thức trọng số của đỉnh Vi .....                            | 17 |
| Hình 2.3 Hệ thống để thực hiện 1 thuật toán xếp hạng dựa trên đồ thị ... | 17 |
| Hình 2.4 Đồ thị vô hướng .....   | 18 |
| Hình 2.5 Công thức thuật toán Textrank trong đồ thị có trọng số.....     | 19 |
| Hình 2.6 Công thức TextRank trích xuất từ khóa .....                     | 23 |
| Hình 2.7 Mô hình sequence-to-sequence với cơ chế attention.....          | 33 |
| Hình 2.8 minh họa ví dụ về chạy thử nghiệm được tác giả công bố.....     | 34 |
| Hình 3.1. Mô hình bài toán tóm tắt văn bản .....                         | 34 |
| Hình 3.2: Thư viện Flask .....   | 35 |
| Hình 3.3: Thư viện Numpy .....   | 37 |
| Hình 3.4: Thư viện Werkzeug .....  | 38 |
| Hình 3.5: Thư viện tempfile .....  | 39 |
| Hình 3.6: Thư viện Underthesea.....                                      | 39 |
| Hình 3.7: Thư viện MoviePy .....   | 41 |
| Hình 3.8: Thư viện SpeechRecognition.....                                | 41 |
| Hình 3.9 Cấu trúc ứng dụng tóm tắt văn bản.....                          | 54 |
| Hình 3.10 Giao diện Login .....  | 55 |
| Hình 3.11 Giao diện Login .....  | 56 |
| Hình 3.12 Giao diện trang chủ.....                                       | 56 |
| Hình 3.13 TH1 Nhập trực tiếp văn bản để test .....                       | 57 |
| Hình 3.14 TH2 Lấy văn bản từ Data để test.....                           | 57 |

|   |    |
|---|----|
| Hình 3.15 TH3 lấy văn bản từ video để test..... | 58 |
| Hình 3.16 Test th1 .....                        | 58 |
| Hình 3.17 Test th2 .....                        | 59 |
| Hình 3.17 Test th3 .....                        | 60 |
| Hình 3.19 Kết quả th1.....                      | 61 |
| Hình 3.20 Kết quả th2.....                      | 62 |
| Hình 3.21 Kết quả th3.....                      | 64 |

## LỜI CẢM ƠN

Trong quá trình thực hiện đồ án tốt nghiệp, với đề tài "xây dựng ứng dụng tóm tắt văn bản tự động bằng ngôn ngữ python", em xin gửi lời cảm ơn chân thành đến tất cả những người đã đóng góp và hỗ trợ hoàn thành thành công đồ án này.

Đầu tiên, em muốn gửi lời cảm ơn sâu sắc tới Giảng viên hướng dẫn, thạc sĩ Ngô Thị Bích Thúy, vì sự chỉ dẫn, hướng dẫn và những kiến thức quý báu mà cô đã truyền đạt cho sinh viên trong suốt quá trình thực hiện đồ án cũng như trong quá trình học. Sự kiên nhẫn và tận tâm của cô đã giúp những sinh viên trong lớp như em vượt qua những khó khăn và hoàn thiện đồ án một cách tốt nhất.

Em cũng muốn bày tỏ lòng biết ơn đến khoa Công nghệ thông tin trường Đại học Công Nghiệp Hà Nội đã tạo điều kiện thuận lợi và cung cấp những kiến thức chuyên môn quan trọng cho em trong suốt quá trình học tập và nghiên cứu tại trường.

Không thể không nhắc đến sự giúp đỡ của các bạn bè cùng lớp trong việc tìm kiếm thông tin, trao đổi ý kiến và hỗ trợ kỹ thuật.

Em cũng muốn bày tỏ lòng biết ơn đến gia đình và người thân đã luôn ủng hộ, động viên và hiểu rõ những khó khăn gấp phai trong quá trình thực hiện đồ án này. Sự động viên và tình yêu thương của gia đình là nguồn động lực quan trọng giúp em vượt qua mọi khó khăn.

Cuối cùng, Em xin gửi lời cảm ơn chân thành đến tất cả những người đã đọc và đánh giá đồ án này. Sự quan tâm và góp ý của mọi người là động lực để em tiếp tục nỗ lực và hoàn thiện hơn trong những nghiên cứu và dự án tương lai.

Xin chân thành cảm ơn!

Sinh viên thực hiện

Khai

Phạm Hữu Khải

## LỜI MỞ ĐẦU

Với sự phát triển mạnh mẽ của công nghệ thông tin và mạng máy tính, lượng tài liệu văn bản khổng lồ được tạo ra với nhiều mục đích sử dụng khác nhau, làm cho việc đọc hiểu và trích lọc thông tin cần thiết trong khối tri thức đồ sộ này trở nên tốn kém về thời gian và chi phí. Điều này trở nên đặc biệt quan trọng với sự gia tăng của các thiết bị cầm tay, đòi hỏi chi phí cho hạ tầng và truyền dẫn thông tin.

Để tăng cường hiệu quả và làm cho việc tiếp nhận thông tin của người dùng trở nên dễ dàng hơn, nhiều nghiên cứu về khai phá dữ liệu và xử lý ngôn ngữ tự nhiên đã được thực hiện. Một trong những lĩnh vực quan trọng nhất trong nghiên cứu này là bài toán tóm tắt văn bản tự động.

Bài toán tóm tắt văn bản tiếng Việt cũng đã được nghiên cứu và áp dụng nhiều kỹ thuật tương tự như trong tiếng Anh. Tuy nhiên, tóm tắt văn bản và xử lý ngôn ngữ tự nhiên cho tiếng Việt đối mặt với nhiều thách thức hơn. Điều này là do tiếng Việt có đặc điểm là tiếng đơn âm và có thanh điệu, khiến cho việc tách từ và tách thành phần ngữ nghĩa trong câu trở nên phức tạp hơn so với tiếng Anh. Hơn nữa, không có nhiều kho dữ liệu tiếng Việt được chuẩn hóa và công bố.

Trong đề tài này, em tập trung nghiên cứu về tóm tắt văn bản tự động theo hướng trích đoạn, sử dụng các mô hình kiến trúc mạng học sâu và các kỹ thuật để xử lý những thách thức trong quá trình tóm tắt văn bản.

**Đô án này được chia thành 3 chương chính:**

**Chương 1:** Trình bày khái quát về bài toán tóm tắt văn bản, bao gồm định nghĩa, phân loại, và các hướng tiếp cận chính trong tóm tắt văn bản, cũng như những đặc trưng và khó khăn khi tóm tắt văn bản tiếng Việt.

**Chương 2:** Mô tả các nghiên cứu liên quan, tập trung vào thuật toán xếp hạng đồ thị như PageRank và TextRank, và cách chúng được áp dụng trong tóm tắt văn bản, đồng thời giới thiệu các phương pháp tóm tắt khác.

**Chương 3:** Trình bày quá trình xây dựng ứng dụng tóm tắt văn bản, bao gồm phát triển mô hình, tiền xử lý dữ liệu, và đánh giá kết quả thực nghiệm, cùng với hạn chế và hướng phát triển của đề tài.

Với đồ án này, em hy vọng không chỉ cung cấp thông tin chi tiết về tóm tắt văn bản tự động và ứng dụng của nó trong ngôn ngữ tiếng Việt, mà còn mở rộng kiến thức và kỹ năng thực tế cho cộng đồng sinh viên và các nhà nghiên cứu. Em kỳ vọng rằng công trình này sẽ góp phần vào sự phát triển của lĩnh vực xử lý ngôn ngữ tự nhiên và khai thác tri thức từ văn bản.

# CHƯƠNG 1: KHÁI QUÁT BÀI TOÁN TÓM TẮT VĂN BẢN

Cùng với sự tăng trưởng mạnh mẽ của mạng Internet, con người ngày càng bị quá tải bởi khối lượng lớn các thông tin và tài liệu trực tuyến. Điều này đã thúc đẩy rất nhiều nghiên cứu về tóm tắt văn bản tự động. Theo Radev và đồng nghiệp, một tóm tắt được định nghĩa như là một văn bản được tạo từ một hoặc nhiều văn bản, truyền đạt các thông tin quan trọng từ các văn bản gốc. Văn bản tóm tắt không dài hơn 50% độ dài văn bản gốc và thông thường bản tóm tắt có độ dài khá ngắn, ngắn hơn nhiều so với 50% độ dài văn bản gốc.

## 1.1 Bài toán tóm tắt văn bản tự động

### 1.1.1 Tóm tắt văn bản

Tóm tắt văn bản tự động là tác vụ để tạo ra một tóm tắt chính xác và hợp ngữ pháp trong khi vẫn giữ được các thông tin chính và ý nghĩa của văn bản gốc. Trong những năm gần đây, có rất nhiều hướng tiếp cận đã được nghiên cứu cho tóm tắt văn bản tự động và đã được áp dụng rộng rãi trong nhiều lĩnh vực. Ví dụ, máy tìm kiếm sinh ra các trích đoạn như là các bản xem trước của tài liệu, các website tin tức sinh ra các đoạn mô tả ngắn gọn cho bài viết (thường là tiêu đề của bài viết).

Mục tiêu của tóm tắt văn bản là tạo ra bản tóm tắt giống như cách con người tóm tắt. Đây là bài toán đầy thách thức, bởi vì khi con người thực hiện tóm tắt một văn bản, chúng ta thường đọc toàn bộ nội dung rồi dựa trên sự hiểu biết và cảm nhận của mình để viết lại một đoạn tóm tắt nhằm làm nổi bật các ý chính của văn bản gốc. Nhưng vì máy tính khó có thể có được tri thức và khả năng ngôn ngữ như của con người, nên việc thực hiện tóm tắt văn bản tự động là một công việc phức tạp.

### 1.1.2. Ứng dụng của tóm tắt văn bản

Tóm tắt văn bản là quá trình rút gọn và trình bày thông tin quan trọng của một đoạn văn bản mà không làm mất đi ý chính của nó. Có nhiều ứng dụng của việc tóm tắt văn bản trong nhiều lĩnh vực khác nhau:

- Tiết kiệm thời gian: Tóm tắt văn bản là một công cụ mạnh mẽ để tiết kiệm thời gian. Trong thế giới ngày nay, khi mọi người đều đối mặt với áp lực thời gian, khả năng nhanh chóng hiểu và tóm tắt nội dung của một đoạn văn bản giúp họ tiết kiệm thời gian quý báu. Điều này đặc biệt quan trọng trong môi trường công việc và học tập, nơi người ta thường xuyên phải xử lý lượng lớn thông tin. Bằng cách này, người đọc có thể nắm bắt ý chính mà không cần phải đọc toàn bộ văn bản, từ đó tối ưu hóa quá trình học và làm việc.
- Hiểu bài đọc: Tóm tắt văn bản không chỉ giúp tiết kiệm thời gian mà còn hỗ trợ quá trình hiểu bài đọc. Khi đọc một đoạn văn bản dài, nguy cơ mất thông tin quan trọng là rất cao. Tóm tắt giúp người đọc tập trung vào những điểm chính, giúp họ hiểu rõ ý chính và ý phụ của văn bản một cách nhanh chóng. Điều này không chỉ làm tăng khả năng hiểu bài đọc mà còn giúp nhớ lâu hơn, vì người đọc đã tập trung vào những điểm quan trọng nhất.
- Nghiên cứu và phân tích: Trong lĩnh vực nghiên cứu, tóm tắt văn bản là một công cụ quan trọng giúp nhà nghiên cứu hiểu nhanh chóng nội dung của các bài báo, công trình nghiên cứu, hoặc báo cáo. Điều này là quan trọng để nắm bắt tình hình nghiên cứu hiện tại và định hình hướng phát triển mới. Nhất là khi có sự cạnh tranh cao trong lĩnh vực nghiên cứu, khả năng nhanh chóng đánh giá và tiếp cận thông tin trở thành chìa khóa cho sự thành công.
- Tìm kiếm thông tin: Trong thế giới kỹ thuật số ngày nay, khi người ta thường xuyên tìm kiếm thông tin trên internet, tóm tắt văn bản trở thành một công cụ quan trọng để nhanh chóng xác định xem nội dung có liên quan hay không. Việc đọc toàn bộ nội dung mỗi trang web có thể là một công việc đầy thách thức và tốn thời gian. Tóm tắt giúp người dùng xác định sự liên quan một cách

nhanh chóng, từ đó tiết kiệm thời gian và công sức trong quá trình tìm kiếm thông tin.

- Quảng cáo và marketing: Trong lĩnh vực quảng cáo và marketing, việc truyền đạt thông điệp một cách ngắn gọn và hiệu quả là chìa khóa để thu hút sự chú ý của khách hàng. Tóm tắt văn bản giúp những thông điệp quảng cáo trở nên dễ tiếp cận và dễ hiểu, tăng khả năng gây ấn tượng và làm tăng hiệu suất chiến lược quảng cáo.
- Chatbot và trí tuệ nhân tạo: Trong lĩnh vực trí tuệ nhân tạo và chatbot, tóm tắt văn bản là yếu tố quan trọng để xây dựng các hệ thống tự động trả lời. Việc này giúp chatbot nhanh chóng trích xuất thông tin quan trọng từ các câu hỏi hoặc yêu cầu của người dùng, cung cấp phản hồi hiệu quả và giải quyết vấn đề một cách nhanh chóng.
- Phân loại và sắp xếp thông tin: Trong lĩnh vực tổ chức thông tin, tóm tắt giúp phân loại và sắp xếp thông tin một cách hiệu quả hơn. Việc này là quan trọng trong quản lý dữ liệu lớn, nơi có hàng nghìn hoặc thậm chí hàng triệu văn bản cần được xử lý. Tóm tắt giúp tạo ra các chỉ mục và danh mục thông tin, giúp người dùng dễ dàng tìm kiếm và truy cập thông tin một cách nhanh chóng.

### 1.1.3. Phân loại tóm tắt văn bản

Tóm tắt văn bản có thể được phân loại theo nhiều tiêu chí khác nhau, tùy thuộc vào mục đích sử dụng và cách thức thực hiện. Dưới đây là một số phân loại phổ biến:

- Theo mức độ tóm tắt:
  - Tóm tắt ngắn (extractive summary): Là việc rút gọn văn bản bằng cách chọn lựa và trích xuất các câu hoặc đoạn văn có ý nghĩa quan trọng mà không tạo thêm nội dung mới.

- **Tóm tắt mở rộng (abstractive summary):** Điều này liên quan đến việc sáng tạo nội dung mới để tóm tắt ý chính của văn bản, thậm chí có thể sử dụng từ ngữ và cấu trúc câu không xuất hiện trong văn bản gốc.

- Theo phương pháp tạo tóm tắt:

- **Tóm tắt tự động (automatic summarization):** Sử dụng các thuật toán máy học và trí tuệ nhân tạo để tạo ra tóm tắt một cách tự động, không cần sự can thiệp của con người.

- **Tóm tắt hướng dẫn (manual summarization):** Đòi hỏi sự can thiệp của con người để đọc và tóm tắt văn bản, có thể thông qua việc đánh dấu câu hoặc đoạn quan trọng.

- Theo phạm vi tóm tắt:

- **Tóm tắt đoạn (sentence summarization):** Tập trung vào việc rút gọn các câu văn bản mà không cần đến cấu trúc đoạn.

- **Tóm tắt đoạn (paragraph summarization):** Hướng tới việc tóm tắt nội dung của các đoạn văn bản.

- Theo mục đích sử dụng:

- **Tóm Tắt Thông Tin (Information Summary):** Tập trung vào việc trích xuất thông tin quan trọng và relevant trong văn bản.

- **Tóm Tắt Ý Chính (Key Point Summary):** Chỉ tập trung vào việc tóm tắt các ý chính và điểm quan trọng của văn bản.

- Theo Ngôn Ngữ và Nền Văn Hóa:

- **Tóm Tắt Ngôn Ngữ Tự Nhiên (Natural Language Summary):** Sử dụng ngôn ngữ tự nhiên để diễn đạt ý chính của văn bản.

- **Tóm Tắt Ngôn Ngữ Kỹ Thuật (Technical Language Summary):** Sử dụng ngôn ngữ chuyên ngành và kỹ thuật để tóm tắt thông tin từ các văn bản chuyên ngành.

- Theo Ứng Dụng Cụ Thể:

- **Tóm Tắt Tin Tức (News Summary):** Tóm tắt thông tin từ các tin tức để cung cấp cái nhìn tổng quan về sự kiện.

- **Tóm Tắt Nghiên Cứu (Research Summary):** Tóm tắt công trình nghiên cứu để giúp nhà nghiên cứu nhanh chóng hiểu về nội dung và kết quả.

#### 1.1.4. Mô hình tóm tắt văn bản

Có nhiều mô hình tóm tắt văn bản khác nhau, từ các phương pháp truyền thống sử dụng kỹ thuật rèn luyện đến các mô hình học máy và học sâu hiện đại. Dưới đây là một số loại mô hình phổ biến:

- Extractive Summarization (tóm tắt khai thác):
  - TextRank: Sử dụng đồ thị để đo lường sự quan trọng của các câu trong văn bản dựa trên mối liên kết giữa chúng. Các câu được xếp hạng và chọn lựa để tạo thành tóm tắt.
  - LexRank: Tương tự như TextRank, LexRank sử dụng mô hình đồ thị để xác định sự quan trọng của các câu.
- Abstractive summarization(tóm tắt trừu tượng):
  - Seq2Seq (Sequence-to-Sequence): Mô hình cơ bản trong abstractive summarization, sử dụng một mô hình encoder-decoder để chuyển đổi văn bản vào một dạng tóm tắt.
  - Pointer-generator networks: Mở rộng Seq2Seq bằng cách cho phép mô hình "chọn" từ vựng từ văn bản gốc thay vì chỉ sử dụng từ vựng đã biết trước.
  - BERT (Bidirectional Encoder Representations From Transformers): Một mô hình học sâu mạnh mẽ, thường được sử dụng cho nhiều nhiệm vụ, có thể được áp dụng cho abstractive summarization thông qua fine-tuning.
- Pre-Trained language models(mô hình ngôn ngữ được đào tạo từ trước):
  - GPT (Generative Pre-trained Transformer): GPT-3, là một trong những mô hình ngôn ngữ tiên tiến nhất, có khả năng thực hiện tóm tắt văn bản thông qua việc sinh ra ngôn ngữ tự nhiên.
  - BERTSUM: Sử dụng mô hình BERT để tổng hợp tóm tắt bằng cách ánh xạ các câu đầu vào vào một không gian vector và tối ưu hóa giá trị tóm tắt.
- Transformer-Based models:

- BART (BART ain't a recursive acronym): Một mô hình transformer được thiết kế đặc biệt cho tác vụ tóm tắt văn bản, có thể thực hiện cả extractive và abstractive summarization.
- T5 (text-to-text transfer transformer): Mô hình transformer đặc biệt được huấn luyện để xử lý tất cả các tác vụ ngôn ngữ tự nhiên dưới dạng "text-to-text," bao gồm cả tóm tắt văn bản.
- Học Tăng Cường (Reinforcement Learning):
  - RLSeq2Seq: Sử dụng học tăng cường để đào tạo mô hình tóm tắt thông qua việc thưởng cho các tóm tắt có chất lượng cao dựa trên sự so sánh với tóm tắt thực tế.
    - Các mô hình này thường được fine-tuned cho nhiệm vụ tóm tắt văn bản trên các tập dữ liệu cụ thể. Sự kết hợp giữa các phương pháp extractive và abstractive cũng được sử dụng để tận dụng những ưu điểm của cả hai loại tóm tắt.

#### 1.1.5. Đánh giá văn bản tóm tắt:

Đánh giá văn bản tóm tắt có thể được thực hiện dựa trên một số tiêu chí quan trọng. Dưới đây là một số điểm chính mà bạn có thể xem xét khi đánh giá một văn bản tóm tắt:

- Chính xác:
  - Extractive Summarization: Đánh giá chính xác của việc chọn lựa và trích xuất câu hoặc đoạn trong tóm tắt. Cần xem xét xem các câu được chọn có bao quát và diễn đạt ý chính của văn bản không.
  - Abstractive Summarization: Đánh giá mức độ chính xác của việc tạo ra tóm tắt mới. Có thể kiểm tra xem mô hình có hiểu và diễn đạt ý chính một cách đầy đủ hay không.
- Chất lượng ngôn ngữ:

- Kiểm tra chất lượng của ngôn ngữ trong tóm tắt. Phân tích ngữ pháp, cú pháp và sự diễn đạt. Có thể sử dụng đánh giá tự động của ngôn ngữ hoặc xem xét của người đánh giá chuyên gia.

- Tính cô đọng:

- Đánh giá xem tóm tắt có thể giữ lại tất cả các ý chính trong một lượng văn bản ngắn không. Cần xem xét xem tóm tắt có giữ được thông tin quan trọng không, hay có sự mất mát nội dung quan trọng.

- Tính tổng quan:

- Xem xét tính tổng quan của tóm tắt. Đánh giá xem tóm tắt có thể tạo ra cái nhìn tổng quan về nội dung của văn bản không. Có thể liên quan đến khả năng của nó trong việc tái tạo tâm trạng, ý đồ, và thông điệp chung của văn bản.

- Khả năng tự động:

- Nếu tóm tắt được tạo ra bởi mô hình tự động, đánh giá sự hiệu quả của mô hình trong việc tạo ra tóm tắt chất lượng. Điều này có thể bao gồm cả đánh giá định lượng và đánh giá chất lượng từ người đánh giá chuyên gia.

- Thực tế và tính ứng dụng:

- Đánh giá về tính thực tế và ứng dụng của tóm tắt. Liên quan đến việc tóm tắt có giúp đỡ hay áp dụng được trong các tình huống thực tế không. Có thể đặt câu hỏi như "Tóm tắt này giúp giải quyết vấn đề gì?" hay "Làm thế nào tóm tắt này hỗ trợ trong công việc hoặc nghiên cứu?"

- Đánh giá so sánh:

- So sánh tóm tắt với văn bản gốc để đảm bảo rằng nó bám sát vào ý chính mà không làm mất thông tin không cần thiết. Đánh giá sự tương đồng giữa tóm tắt và văn bản nguồn.

- Đánh giá dựa trên điểm số:

- Sử dụng hệ thống điểm số để đánh giá từng tiêu chí. Điều này có thể bao gồm việc xác định trọng số cho mỗi tiêu chí dựa trên độ quan trọng tương ứng.

- Phản hồi người dùng:

- Thu thập phản hồi từ người đọc hoặc người sử dụng tóm tắt. Có thể thực hiện thông qua cuộc khảo sát, đánh giá trực tuyến, hoặc các phương tiện khác để đo lường sự hài lòng và hiệu suất.

- Tính tương thích:

- Đánh giá tính tương thích của tóm tắt với các nền tảng hay ứng dụng cụ thể, đặc biệt nếu tóm tắt được tạo ra để tích hợp vào các hệ thống hoặc ứng dụng khác. Điều này có thể bao gồm khả năng tích hợp với API hoặc hệ thống khác.

#### 1.1.6. Một số đặc trưng và khó khăn trong tóm tắt văn bản tiếng Việt

Tóm tắt văn bản tiếng Việt đối mặt với nhiều đặc trưng và thách thức độc đáo do ngôn ngữ phong phú và đặc sắc của tiếng Việt. Mỗi khía cạnh của việc tóm tắt đều đặt ra những yêu cầu và vấn đề cụ thể:

- Trước hết, sự phong phú về từ ngữ và ngữ pháp trong tiếng Việt là một điểm đặc trưng. Việc lựa chọn câu và đoạn văn để tạo ra tóm tắt chính xác đòi hỏi sự hiểu biết sâu sắc về ngôn ngữ và cấu trúc ngữ cảnh.
- Sự tự lập của các câu tiếng Việt là một thách thức khác. Các câu thường đứng độc lập và không có sự liên kết rõ ràng, điều này có thể làm tăng độ phức tạp trong việc xây dựng một tóm tắt có tính nhất quán.
- Sự xuất hiện của các yếu tố nghệ thuật và tả tưởng trong văn bản tiếng Việt làm tăng độ khó khăn trong việc diễn đạt ý chính một cách chính xác trong tóm tắt. Các mô hình tóm tắt cần phải hiểu và tái tạo được sự nghệ thuật và tả tưởng này.
- Đa dạng văn hóa và ngôn ngữ giữa các khu vực và cộng đồng sử dụng tiếng Việt là một đặc điểm độc đáo khác. Điều này đòi hỏi mô hình tóm tắt phải đổi mới với sự đa dạng ngôn ngữ và ngữ cảnh.
- Đối với dữ liệu tiếng Việt, sự đồng nhất không luôn được đảm bảo. Sự đa dạng về cách diễn đạt ý chính và cấu trúc văn bản có thể làm tăng thách thức trong việc huấn luyện mô hình chung cho tóm tắt.

- Cuối cùng, từ vựng chuyên ngành và thuật ngữ xuất hiện thường xuyên, đặc biệt là trong các lĩnh vực chuyên sâu. Điều này đặt ra yêu cầu cao về sự hiểu biết về ngôn ngữ chuyên ngành và khả năng tạo ra tóm tắt phù hợp.
- Những thách thức này đều đòi hỏi sự linh hoạt và hiểu biết sâu sắc về ngôn ngữ và văn hóa tiếng Việt khi phát triển các mô hình tóm tắt hiệu quả.

### 1.2. Các hướng tiếp cận tóm tắt văn bản

Nhìn chung, có hai hướng tiếp cận chính cho bài toán tóm tắt văn bản tự động là trích chọn (extraction) và tóm lược (abstraction). Tóm tắt văn bản có thể được phân loại dựa trên đầu vào (đơn hoặc đa văn bản), mục đích (tổng quát, theo lĩnh vực cụ thể, hoặc dựa trên truy vấn) và loại đầu ra (trích chọn hoặc tóm lược).

Phương pháp tóm tắt trích chọn thực hiện đánh giá các phần quan trọng của văn bản và đưa chúng một cách nguyên bản vào bản tóm tắt. Do đó, phương pháp này chỉ phụ thuộc vào việc trích chọn các câu từ văn bản gốc dựa trên việc xếp hạng mức độ liên quan của các cụm từ để chỉ chọn những cụm từ liên quan nhất tới nội dung của tài liệu gốc.

Trong khi đó, phương pháp tóm tắt tóm lược nhằm tạo ra văn bản tóm tắt mới có thể không bao gồm các từ hoặc cụm từ trong văn bản gốc. Nó cố gắng hiểu và đánh giá văn bản sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến để tạo ra một văn bản ngắn hơn, truyền đạt được những thông tin quan trọng nhất từ văn bản gốc.

Mặc dù các tóm tắt được con người thực hiện thường không giống như trích chọn, tuy nhiên, hầu hết các nghiên cứu về tóm tắt văn bản hiện tại vẫn tập trung vào tóm tắt bằng phương pháp trích chọn. Điều này là bởi vì, về cơ bản, các tóm tắt sinh ra bởi phương pháp trích chọn thường cho kết quả tốt hơn so với tóm tắt bằng phương pháp tóm lược. Điều này là do phương pháp tóm tắt bằng tóm lược phải đổi mới với các vấn đề như thể hiện ngữ nghĩa,

suy luận và sinh ngôn ngữ tự nhiên, những vấn đề này phức tạp hơn nhiều so với việc trích chọn câu. Hướng tiếp cận tóm tắt bằng tóm lược khó hơn so với tóm tắt bằng trích chọn, nhưng phương pháp này được kỳ vọng có thể tạo ra các văn bản tóm tắt giống như cách con người thực hiện.

## CHƯƠNG 2: MỘT SỐ NGHIÊN CỨU VỀ TÓM TẮT VĂN BẢN

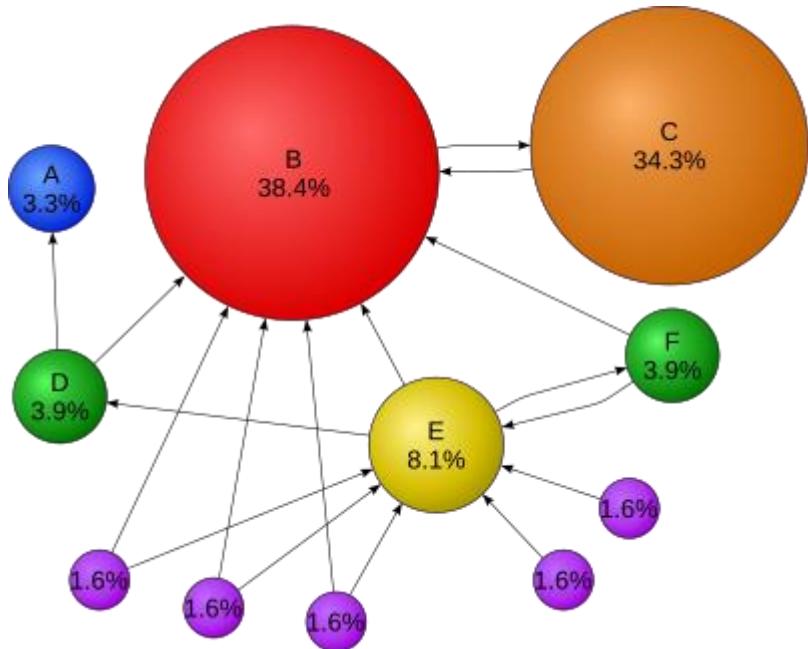
### 2.1 Tổng quan về tóm tắt nội dung

Tóm tắt nội dung là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (NLP) và khai thác tri thức, với mục tiêu rút gọn và trình bày các thông tin cốt lõi từ một nguồn dữ liệu ban đầu mà không làm mất đi các ý chính. Sự phát triển nhanh chóng của thông tin và nội dung số, đặc biệt là văn bản và video, đã dẫn đến nhu cầu ngày càng cao về việc tóm tắt tự động.

Trong tóm tắt nội dung, có hai phương pháp tiếp cận chính: tóm tắt trích chọn (extractive summarization) và tóm tắt tóm lược (abstractive summarization).

Tóm tắt trích chọn: Phương pháp này tập trung vào việc chọn lọc các câu hoặc đoạn văn có ý nghĩa quan trọng từ văn bản gốc. Các câu này sau đó được ghép lại thành một bản tóm tắt mà không thay đổi cấu trúc hay ngữ nghĩa ban đầu. Đây là cách tiếp cận phổ biến do tính đơn giản và hiệu quả của nó, đặc biệt trong các hệ thống tóm tắt tự động dựa trên quy tắc.

Tóm tắt tóm lược: Khác với tóm tắt trích chọn, phương pháp này tạo ra một bản tóm tắt mới bằng cách diễn giải lại nội dung từ văn bản gốc. Hệ thống sẽ hiểu ngữ nghĩa của văn bản và viết lại thông tin theo một cách khác, có thể là ngắn gọn hơn và súc tích hơn so với văn bản gốc. Tóm tắt tóm lược phức tạp hơn về mặt kỹ thuật, nhưng nó có tiềm năng cung cấp những bản tóm tắt chất lượng hơn, gần gũi với cách viết của con người.



Hình 2.1 Thuật toán PageRank

## 2.2. Phương pháp tóm tắt văn bản

Trong các nghiên cứu và ứng dụng trước đây, tóm tắt văn bản trích chọn đã được áp dụng rộng rãi với các kỹ thuật khác nhau để chọn ra các câu quan trọng nhất từ văn bản gốc. Các kỹ thuật chính bao gồm:

- Tính toán độ tương đồng bằng Cosine Similarity:

Cosine Similarity là một kỹ thuật đo lường mức độ tương đồng giữa hai vector bằng cách tính cosin của góc giữa chúng. Giá trị cosin gần với 1 cho thấy hai vector (câu) có nội dung tương đồng cao, trong khi giá trị gần 0 chỉ ra sự khác biệt lớn.

Trong quy trình của bạn, Cosine Similarity được sử dụng để đánh giá mức độ tương đồng giữa các câu trong văn bản, từ đó xác định những câu nào có nội dung tương tự và loại bỏ các câu trùng lặp hoặc ít quan trọng hơn.

- Thuật toán TextRank:

TextRank là một thuật toán dựa trên đồ thị, phát triển từ thuật toán PageRank, được sử dụng rộng rãi để tóm tắt văn bản bằng cách

xếp hạng các câu trong một văn bản dựa trên mức độ quan trọng của chúng.

TextRank tạo ra một đồ thị, trong đó mỗi đỉnh là một câu và các cạnh giữa các đỉnh biểu diễn mối quan hệ hoặc sự tương đồng giữa các câu. Trọng số của các cạnh được xác định bởi Cosine Similarity giữa các vector câu. Sau đó, thuật toán tiến hành tính toán để xếp hạng các câu, từ đó chọn ra các câu có trọng số cao nhất để tạo thành bản tóm tắt.

### 2.2.1 Mô hình TextRank

Thuật toán xếp hạng dựa trên đồ thị là cách đưa ra cách chọn đỉnh quan trọng trong đồ thị dựa trên các thông tin toàn cục của các đỉnh trong đồ thị. Ý tưởng của thuật toán này dựa trên hai yếu tố: bỏ phiếu và đề cử. ". Khi đỉnh đầu tiên liên kết với đỉnh thứ hai, ví dụ như thông qua mối quan hệ kết nối hoặc cạnh biểu đồ. Mỗi một liên kết đến đỉnh đang xét thì nó được 1 phiếu bầu. Như vậy, càng nhiều phiếu bầu thì đỉnh đó càng quan trọng. Từ cách xác định trên thì trọng số của một đỉnh chính là số phiếu bầu cho đỉnh đó.

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

$V$ : là tập các đỉnh

$E$ : là tập các cạnh của đồ thị,  $E$  là tập con của  $V \times V$  ( $E \subseteq V \times V$ ). Với mỗi đỉnh  $Vi$  thì ta có:

-  $In(V_i)$  là tập các đỉnh trỏ đến  $Vi$

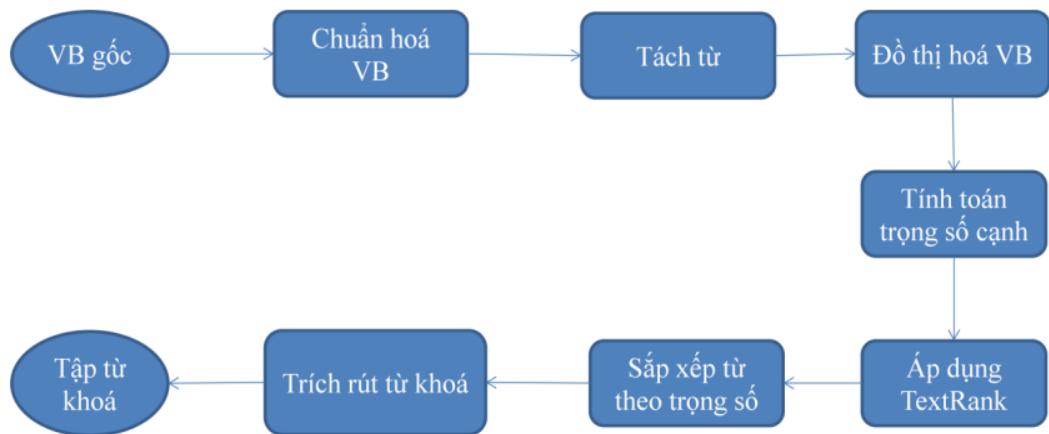
-  $Out(V_i)$  là tập các đỉnh mà  $Vi$  trỏ đến.

Trọng số của đỉnh  $Vi$  được xác định như sau: (Brin and Page, 1998):

$$S(V_j) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (1)$$

Hình 2.2 Công thức trọng số của đỉnh Vi

Trong đó  $d$  là nhân tố giảm, có giá trị từ 0 đến 1. Nó là xác suất mà một đỉnh có liên kết đến một đỉnh bất kỳ trong đồ thị. Đối với các trang web thì  $d$  là xác suất người dùng nhấp vào một liên kết bất kỳ và xác suất để người dùng vào một trang web hoàn toàn mới là  $1 - d$ . Theo PageRank thì  $d = 0.85$ . Đây cũng là xác suất sẽ được sử dụng trong TextRank. Ban đầu gán cho tất cả các đỉnh trong đồ thị các giá trị khởi tạo và tính toán lặp lại cho đến khi kết quả hội tụ lại đạt ngưỡng xác định. Sau quá trình tính toán thì trọng số của mỗi đỉnh chính là mức độ quan trọng của đỉnh đó trong toàn đồ thị. Có điều cần lưu ý, đó là giá trị trọng số của mỗi đỉnh sẽ không phụ thuộc vào giá trị khởi tạo ban đầu được gán cho mỗi đỉnh. Ngoài ra thì số lượng 16 các vòng lặp tính toán để ra được trọng số là khác nhau. Để hiểu rõ thuật toán hơn ta có hình vẽ sau:

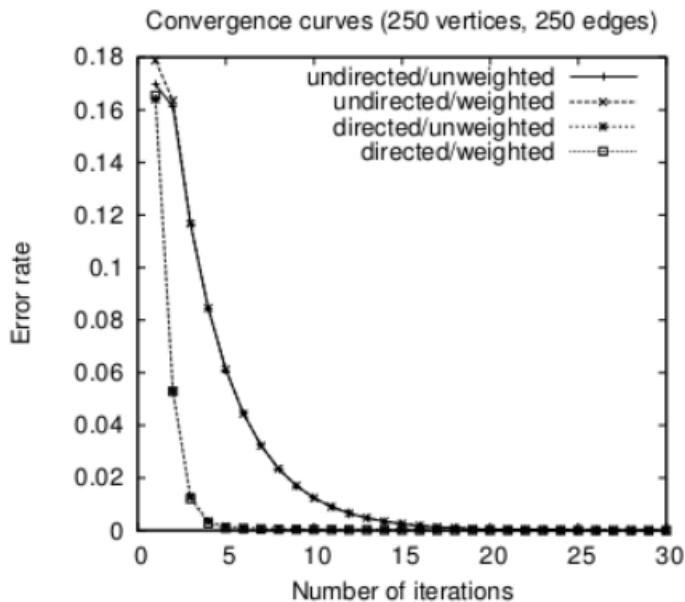


Hình 2.3 Hệ thống để thực hiện 1 thuật toán xếp hạng dựa trên đồ thị

## 2.2.2 Đồ thị vô hướng

Việc áp dụng thuật toán TextRank vào đồ thị vô hướng cũng giống như với đồ thị có hướng. Có một điểm cần lưu ý, đó là trong đồ thị vô hướng thì số đỉnh vào bằng số đỉnh ra.

Ta có các hình vẽ sau:



Hình 2.4 Đồ thị vô hướng

Trong hình 10 thì đường cong hội tụ cho đồ thị được sinh ngẫu nhiên với 250 đỉnh và 250 cạnh, với ngưỡng dừng là 10-5 (ngưỡng này được xác định đủ 17 nhở để thuật toán dừng tính toán) cho thấy số lần lặp của quá trình tính toán không cao mặc dù số lượng đỉnh và cạnh lớn. Bên cạnh đó thì đường cong độ tụ của đồ thị có hướng và vô hướng gần như trùng nhau. Điều đó cho thấy đồ thị vô hướng hay có hướng đều có kết quả giống nhau, chỉ khác nhau ở số lần tính toán lặp lại.

### 2.2.3 Đồ thị có trọng số

Vì thuật toán PageRank ban đầu chỉ sử dụng đồ thị không trọng số do gần như không có tình huống một trang web có nhiều liên kết đến một trang nào đó trong môi trường web. Tuy nhiên đổi với các văn bản trong ngôn ngữ tự nhiên thì việc một văn bản nào đó có nhiều thành phần tham chiếu đến một văn bản khác là hoàn toàn xảy ra. Do đó, để cải tiến PageRank cho phù hợp

với ngôn ngữ tự nhiên, thuật toán Textrank sử dụng đồ thị có trọng số. Trọng số ở đây được định nghĩa là độ dài kết nối giữa hai đỉnh Vi và Vj kí hiệu  $w_{ij}$ . Từ đó suy ra công thức (1) phải được thay đổi để phù hợp với đồ thị có trọng số trong thuật toán Textrank. Ta được công thức mới như sau:

$$S(V_j) = (1 - d) + d * \sum_{j \in In(V_j)} \frac{w_{ij}}{\sum_{v_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (2)$$

Hình 2.5 Công thức thuật toán Textrank trong đồ thị có trọng số

Như vậy, theo hình (1) ở trên thì số lần lặp lại tính toán để có độ tụ đạt ngưỡng 10-5 của đồ thị có trọng số và đồ thị không có trọng số là tương đương nhau.

#### 2.2.4 Đồ thị hóa văn bản

Hiện nay, trên thế giới có một số công trình xử lý văn bản sử dụng mô hình đồ thị. Các mô hình đồ thị tương đối đa dạng và mỗi mô hình mang nét đặc trưng riêng. Mỗi đồ thị là một văn bản hoặc biểu diễn cho tập văn bản. Đỉnh của đồ thị có thể là câu, hoặc từ, hoặc kết hợp câu và từ. Cạnh nối giữa các đỉnh là vô hướng hoặc có hướng, thể hiện mối quan hệ trong đồ thị. Nhãn đỉnh thường là tần số xuất hiện của đỉnh. Còn nhãn cạnh là tên mối liên kết khái niệm giữa 2 đỉnh, hay tần số xuất hiện chung của 2 đỉnh trong một phạm vi nào đó, hay tên vùng mà đỉnh xuất hiện. Trong bài toán trích rút 18 từ khoá, thì đỉnh là từ, cạnh thể hiện sự tương đồng giữa các từ. Do từ lưu giữ được nhiều thông tin cấu trúc nhất nên mô hình đồ thị sử dụng đỉnh là từ được nghiên cứu sâu hơn và có nhiều biến thể nhất. Ưu điểm của mô hình đồ thị sử dụng đỉnh là từ trong văn bản là mô hình hoá văn bản một cách trực quan, logic, thể hiện được quan hệ ngữ nghĩa giữa các khái niệm và cho kết quả truy vấn thông tin chính xác hơn[5]. Văn bản trên web là một chuỗi các ký tự / từ được sắp xếp với nhau. Vậy nên để áp dụng được vào thuật toán dùng đồ thị để đại diện cho văn bản, các liên kết giữa các từ, cụm từ, câu hoặc các quan

hệ ngữ nghĩa. Tuỳ thuộc vào các ứng dụng mà kích thước văn bản, các đặc trưng được đưa vào đồ thị là từ, cụm từ, hay cả câu. Cũng giống như việc xác định các đỉnh trong đồ thị như trên thì việc xác định các cạnh trong đồ thị là gì cũng phụ thuộc vào miền ứng dụng. Quan hệ được xác định có thể là từ vựng, ngữ nghĩa hoặc ngữ cảnh.

Tuỳ vào các loại và đặc trưng để đưa vào đồ thị mà có các cách thức làm việc. nhưng cách thức hoạt động của thuật toán xếp hạng dựa trên đồ thị áp dụng cho ngôn ngữ tự nhiên có các bước như sau:

- Xác định đơn vị văn bản dùng tốt nhất cho từng công việc, thêm vào là đỉnh của đồ thị.
- Xác định quan hệ kết nối giữa các đơn vị văn bản đã xác định ở trên để vẽ các cạnh giữa các đỉnh trong đồ thị. Các cạnh này có thể là vô hướng hoặc có hướng, có trọng số hoặc không có trọng số
- Lặp lại thuật toán xếp hạng cho đến khi độ tụ thoả mãn ngưỡng.
- Sắp xếp các đỉnh dựa trên các trọng số đã được tính toán trong bước trên. Như vậy, thuật toán này giúp cho em làm được hai việc: Trích rút từ khoá và trích rút câu trong văn bản ngôn ngữ tự nhiên. Vấn đề được đề cập ngay sau đây.

#### 2.2.5 Sử dụng TextRank trích xuất từ khóa

Năm 2003, Hulth đã dùng hệ thống học máy giám sát để trích xuất từ khoá kết hợp cả các đặc trưng về từ vựng và cú pháp. Trong nghiên cứu của mình, Hulth chỉ sử dụng bản tóm tắt để trích rút từ khoá thay vì toàn văn bản vì theo bà, văn bản trên Internet tồn tại chủ yếu ở dạng tóm lược. Đối với thuật toán TextRank, việc trích rút từ khoá cũng được thực hiện đối với văn bản tóm lược. Mặc dù vậy thì việc áp dụng cho toàn văn bản là hoàn toàn khả thi.[6]

Mục đích của việc trích xuất từ khoá tự động là tìm ra các cụm từ mô tả văn bản tốt nhất. Các từ khoá này có thể dùng cho nhiều mục đích khác nhau như phân lớp văn bản hay tóm tắt văn bản tự động. Trong các cách để trích xuất từ khoá thì cách trích xuất các từ khoá có tần suất xuất hiện nhiều nhất là dễ nhất. Mặc dù vậy thì kết quả của phương pháp này không tốt. Điều này đã thúc đẩy các nhà khoa học tìm ra các phương pháp khác hiệu quả hơn. Trong số đó có phương pháp sử dụng học máy có giám sát để trích xuất từ khoá dựa trên các đặc trưng về từ vựng và cú pháp. Phương pháp này lần đầu tiên được biết đến vào năm 1999, trong đó việc kết hợp tham số hoá các nguyên tắc phỏng đoán và thuật toán di truyền vào hệ thống rút từ khoá sẽ tự động nhận dạng các từ khoá trong tài liệu. Một thuật toán khác cũng được đưa ra trong năm 1999 sử dụng phương pháp học máy Naïve Bayes đã nâng cao chất lượng từ khoá trích rút được.

Đơn vị để xếp hạng trong thuật toán TextRank đối với quá trình trích rút từ khoá là chuỗi của một hoặc nhiều từ vựng được rút ra từ văn bản và chúng là các đỉnh trong đồ thị. Bất kỳ quan hệ nào nào giữa 2 đơn vị từ vựng hữu ích cho việc đánh giá thì đều được thêm vào là cạnh của đồ thị. Ở đây ta sử dụng quan hệ đồng xuất hiện, nó được xác định bởi khoảng cách giữa các từ đồng xuất hiện trong văn bản; hai đỉnh được xác định là nối với nhau khi khoảng cách đồng xuất hiện của hai đơn vị từ vựng không quá  $N$  từ với  $2 \leq N \leq 10$ . Các liên kết đồng xuất hiện thể hiện mối quan hệ giữa các yếu tố cú pháp, nó cũng tương tự như các liên kết ngữ nghĩa để tìm ra từ có nghĩa nhập nhằng, chúng đại diện cho các chỉ số của một văn bản.

Các đỉnh được thêm vào đồ thị bị giới hạn bởi các bộ lọc ngữ nghĩa, nó chỉ chọn các đơn vị từ vựng phù hợp, ví dụ như chọn danh từ, động từ và tạo các cạnh nối giữa các danh từ và động từ đó. Từ đó, ta tạo ra nhiều bộ lọc ngữ nghĩa để cho kết quả tốt hơn.

Thuật toán trích rút từ khoá TextRank là thuật toán hoàn toàn không giám sát. Cách thức hoạt động như sau:

Tách từ và gán nhãn, có các bộ lọc ngữ nghĩa. Để tránh gia tăng kích thước đồ thị thì áp dụng các đơn vị từ vựng phái có độ dài nhất định( n- gram).

Đưa tất cả các đơn vị từ vựng có ở bước trên vào đồ thị. Các cạnh được đưa vào để liên kết các đơn vị từ vựng đồng xuất hiện với khoảng cách N từ. Sau khi dựng xong đồ thị( vô hướng, không trọng số) thì khởi tạo trọng số cho các đỉnh giá trị là 1. Và theo hình 10 thì số lần lặp lại từ 20- 30 của thuật toán sẽ cho kết quả đạt ngưỡng 10-5 . Sau khi có kết quả cho mỗi đỉnh thì thực hiện quá trình sắp xếp ngược trọng số. T đỉnh đầu tiên sẽ được đưa vào quá trình tiếp theo,  $5 \leq T \leq 20$ . Ở đây thì T được lấy theo kích thước văn bản đầu vào.

Sau bước trên ta được một tập các đơn vị từ vựng. Các đơn vị liền kề nhau thì được ghép lại với nhau để tạo thành từ khoá dài. ↗ Thuật toán TextRank gồm 5 giai đoạn như sau:

Bước 1:

Phần xử lý ngôn ngữ tự nhiên sử dụng thuật toán của Stanford (open source). Kết quả trả về là một tập các terms. Một term có thể là một danh từ, hoặc một tính từ

Ví dụ: trong câu: “the cars are loaded onto a train car with the help of Wrench” thì các term là: cars| train| car| help|Wrench.

Bước 2:

Tiếp theo sử dụng thuật toán TextRank để đánh trọng số cho các term trong bước 1. Ý tưởng là như sau: ( Theo bài báo của Rada Mihalcea and Paul Tarau, 2004)

Tất cả các term sẽ được biểu diễn như các đỉnh của graph, 2 term được nối với nhau nếu chúng cùng thuộc một sentence và cách nhau từ 2 terms.- 10 terms

Ví dụ: Từ các term ở trên thì cars sẽ được liên kết với train, car. Term train sẽ được liên kết với các term cars, car, help. 21

Như vậy một graph đã được xây dựng. Để đánh trọng số cho các đỉnh của graph, em sử dụng thuật toán được phát triển từ thuật toán PageRank trong bài báo mới nhất

Giả sử đối với mỗi đỉnh , gọi là trọng số của nó. Vậy thì phương trình quan hệ giữa đỉnh và các đỉnh kề của nó sẽ là:

$$S(v_i) = (1 - d) + d \times \sum_{v_j \in C(v_i)} \frac{attr(v_i, v_j)}{\sum_{v_k \in C(v_j)} attr(v_j, v_k)} \times S(v_j)$$

Hình 2.6 Công thức TextRank trích xuất từ khóa

$$attr(v_i, v_j) = \frac{freq(v_i) \times freq(v_j)}{freq(v_i) + freq(v_j)}$$

Trong đó  $d = 0.85$  là hằng số của thuật toán

$$freq(v_j)$$

ở đó  $freq(v_i)$  là tần số xuất hiện của từ trong văn bản là tần số xuất hiện của từ trong văn bản

- Giải hệ thống phương trình hàm này bằng cách đưa vào các giá trị trong khởi tạo bất kỳ và số vòng lặp, chúng ta đạt được các trọng số cho mỗi đỉnh
- Sau bước b) chúng ta lấy ra 5% các đỉnh có giá trị trọng số cao nhất. Một đỉnh có trọng số càng cao nếu như đỉnh đó xuất hiện nhiều lần trong văn bản hoặc có nhiều liên kết đến các đỉnh khác hoặc có liên kết đến các đỉnh có trọng số cao khác.

- Chúng ta coi các đỉnh này sẽ là các topic chính của phim.

Bước 4: Sử dụng thuật toán n-gram để tìm các keyword phrase từ các term tìm được trong bước 1. Trọng số của phrase sẽ bằng tổng các trọng số của các term mà nó chứa được tính trong bước 3.

### 2.2.6 Giải thuật TextRank

TextRank là một kỹ thuật tóm tắt văn bản theo phương pháp extractive và trong học máy thì là học không giám sát (Unsupervised Learning). TextRank không dựa trên bất kỳ dữ liệu đào tạo nào trước đó và có thể hoạt động với bất kỳ đoạn văn bản tùy ý nào.

Từ khóa là một từ hay một cụm từ dùng để mô tả một cách chính xác, ngắn gọn nhất nội dung chính của một tài liệu (văn bản, hay các trang web). Trong tiếng Anh, từ khóa được thể hiện dưới nhiều thuật ngữ khác nhau như: keywords, term, query term, hay tags; nhưng ý nghĩa của chúng là giống nhau.

Về cơ bản các bước của tiến trình trích rút thông tin như sau:

Theo tiến sĩ Diana Maynard, hầu hết các hệ thống trích rút thông tin nói chung thường tiến hành các bước sau:

Tiền xử lý:

- Nhận biết định dạng tài liệu( Format detection)
- Tách từ ( Tokenization)
- Phân đoạn từ( Word segmentation)
- Giải quyết nhập nhằng ngữ nghĩa( Sense disambiguation)
- Tách câu( Sentence splitting)
- Gán nhãn từ loại( POS tagging)

Phương pháp bao gồm việc xác định một số đơn vị văn bản dựa trên văn bản ngôn ngữ tự nhiên, kết hợp nhiều đơn vị văn bản với nhiều nút biểu

đò, và xác định ít nhất một mối quan hệ kết nối giữa ít nhất hai trong số nhiều đơn vị văn bản. Phương pháp này cũng bao gồm liên kết ít nhất một mối quan hệ kết nối với ít nhất một cạnh biểu đồ kết nối ít nhất hai trong số nhiều nút biểu đồ và xác định nhiều thứ hạng liên quan đến nhiều nút biểu đồ dựa trên ít nhất một cạnh biểu đồ.

### 2.2.7 Các phương pháp tính độ tương đồng

Trong giải thuật TextRank, độ tương đồng thường được sử dụng để xác định mối quan hệ giữa các từ hoặc câu trong văn bản. Độ tương đồng giúp đo lường mức độ tương tự giữa các phần của văn bản dựa trên các đặc trưng như tần suất xuất hiện cùng nhau, ngữ cảnh chung, hoặc các đặc trưng ngữ nghĩa khác.

Các phương pháp chi tiết hơn để tính độ tương đồng giữa các từ hoặc câu trong giải thuật TextRank:

- Tần suất xuất hiện cùng nhau (Co-occurrence): Tính số lần mà hai từ xuất hiện cùng nhau trong một cửa sổ trượt qua văn bản. Xác định "ngữ cảnh" bằng cách chọn một số lượng từ xung quanh mỗi từ để xem xét.
- Ngữ cảnh chung (Context Overlap): Đo lường sự giống nhau giữa các ngữ cảnh của hai từ hoặc câu. Sử dụng các đặc trưng ngữ cảnh chung như số từ chung, số câu chung, hay mô hình hóa ngữ cảnh như một vector.
- Độ đo Tần số Từ - Nghịch Đảo Tần Số Tài Liệu (TF-IDF): Tính TF-IDF cho từng từ trong văn bản. Sử dụng công thức cosine similarity giữa hai vectơ TF-IDF để tính độ tương đồng.
- Jaccard Similarity: Cho tập hợp các từ hoặc câu, tính Jaccard similarity bằng cách lấy kích thước của giao của hai tập hợp chia cho kích thước của hợp của chúng.

- Dice Similarity: Tính giống nhau Dice giữa hai tập hợp bằng cách sử dụng công thức  $2 * (|A \cap B|) / (|A| + |B|)$ , nơi A và B là hai tập hợp.
- Overlapping Coefficient: Tính hệ số chồng chéo giữa hai tập hợp bằng cách sử dụng công thức  $|A \cap B| / \min(|A|, |B|)$ .

Các phương pháp này có thể được lựa chọn dựa trên đặc điểm của văn bản cụ thể và mục tiêu cụ thể của ứng dụng TextRank. Việc chọn phương pháp phù hợp có thể cải thiện hiệu suất của hệ thống trích xuất thông tin hoặc tóm tắt văn bản. Pháp này có thể được lựa chọn dựa trên đặc điểm của văn bản cụ thể và mục tiêu cụ thể của ứng dụng TextRank.

## 2.3 Phương pháp tóm tắt video

Với sự phát triển mạnh mẽ của công nghệ và nhu cầu tiêu thụ nội dung số, tóm tắt video đang trở thành một lĩnh vực nghiên cứu quan trọng, đặc biệt khi lượng thông tin từ video ngày càng tăng. Quá trình tóm tắt video có thể được thực hiện qua các bước sau:

### 2.3.1. Chuyển đổi video sang văn bản

- Tách âm thanh từ video: Bước đầu tiên trong quá trình tóm tắt video là tách phần âm thanh ra khỏi video. Điều này được thực hiện bằng cách sử dụng thư viện moviepy, một công cụ mạnh mẽ cho việc xử lý video và âm thanh trong Python.
- Nhận diện giọng nói (Speech Recognition): Sau khi tách âm thanh, bước tiếp theo là chuyển đổi âm thanh này thành văn bản. Thư viện speech\_recognition trong Python được sử dụng để thực hiện quá trình này, cho phép hệ thống nhận diện và chuyển đổi giọng nói thành văn bản. Đây là một bước quan trọng, bởi nó tạo ra dữ liệu đầu vào cho quá trình tóm tắt tương tự như tóm tắt văn bản.

### 2.3.2. Xử lý văn bản đã chuyển đổi

Sau khi có văn bản từ video, các kỹ thuật xử lý văn bản mà bạn đã áp dụng cho tóm tắt văn bản có thể được tái sử dụng:

- Tiền xử lý: Văn bản thu được từ nhận diện giọng nói thường chứa nhiều lỗi phát âm, ngữ pháp hoặc từ ngữ không cần thiết. Quá trình tiền xử lý giúp làm sạch và chuẩn hóa văn bản trước khi tiến hành các bước tiếp theo.
- Biểu diễn và tính toán độ tương đồng: Giống như tóm tắt văn bản, văn bản từ video cũng được biểu diễn dưới dạng các ma trận, và độ tương đồng giữa các câu được tính toán bằng Cosine Similarity.

- Áp dụng TextRank: Thuật toán TextRank được sử dụng để xếp hạng các câu từ văn bản chuyển đổi từ video, sau đó chọn ra các câu có trọng số cao nhất để tạo thành bản tóm tắt.

#### 2.4. Tóm tắt văn bản theo hướng trích chọn

Như đã đề cập trong chương 1, các kỹ thuật tóm tắt bằng trích chọn sinh ra các đoạn tóm tắt bằng cách chọn một tập các câu trong văn bản gốc. Các đoạn tóm tắt này chứa các câu quan trọng nhất của đầu vào. Đầu vào có thể là đơn văn bản hoặc đa văn bản. Trong khuôn khổ của đề tài này, đầu vào của bài toán tóm tắt văn bản là đơn văn bản.

Các hệ thống tóm tắt văn bản theo hướng trích chọn thường gồm các tác vụ sau: xây dựng một đại diện trung gian (intermediate representation) của văn bản đầu vào thể hiện các đặc điểm chính của văn bản; tính điểm (xếp hạng) các câu dựa trên đại diện trung gian đã xây dựng; chọn các câu để đưa vào tóm tắt.

Mỗi hệ thống tóm tắt văn bản tạo ra một số đại diện trung gian của văn bản mà nó sẽ thực hiện tóm tắt và tìm các nội dung nổi bật dựa trên đại diện trung gian này. Có hai hướng tiếp cận dựa trên đại diện trung gian là chủ đề đại diện (topic representation) và các đặc trưng đại diện (indicator representation).

Các phương pháp dựa trên chủ đề đại diện biến đổi văn bản đầu vào thành một đại diện trung gian và tìm kiếm các chủ đề được thảo luận trong văn bản. Kỹ thuật tóm tắt dựa trên chủ đề đại diện tiêu biểu là phương pháp tiếp cận dựa trên tần suất (frequency). Phương pháp dựa trên các đặc trưng đại diện mô tả các câu trong văn bản như một danh sách các đặc trưng quan trọng, chẳng hạn như độ dài câu, vị trí của câu trong tài liệu, hoặc câu có chứa những cụm từ nhất định.

Khi các đại diện trung gian đã được tạo ra, một điểm số thể hiện mức độ quan trọng sẽ được gán cho mỗi câu. Đối với phương pháp dựa trên chủ đề đại diện, điểm số của một câu thể hiện mức độ giải thích của câu đối với một vài chủ đề quan trọng nhất của văn bản. Trong hầu hết các phương pháp dựa trên đặc trưng đại diện, điểm số được tính bằng tổng hợp các dấu hiệu từ các đặc trưng khác nhau. Các kỹ thuật học máy thường được sử dụng để tìm trọng số cho các đặc trưng.

Cuối cùng, hệ thống tóm tắt sẽ lựa chọn các câu quan trọng nhất để tạo ra bản tóm tắt. Có thể áp dụng các thuật toán tham lam để chọn các câu quan trọng nhất từ văn bản gốc, hoặc biến việc lựa chọn câu thành một bài toán tối ưu trong đó xem xét ràng buộc tối đa hóa tầm quan trọng tổng thể và sự gắn kết ngữ nghĩa trong khi tối thiểu hóa sự đụng độ. Có nhiều yếu tố khác cần được cân nhắc khi lựa chọn các câu quan trọng, ví dụ như ngữ cảnh của bản tóm tắt hay loại tài liệu cần tóm tắt (bài báo tin tức, email, báo cáo khoa học). Các tiêu chí này có thể trở thành các trọng số bổ sung cho việc lựa chọn các câu quan trọng để đưa vào bản tóm tắt.

## 2.4 Tóm tắt văn bản theo hướng chính chọn

### 2.4.1. Phương pháp chủ đề đại diện dựa trên tần xuất

- Xác Suất của Từ (Word Probability)

Xác suất của từ (word probability) là dạng đơn giản nhất sử dụng tần xuất trên văn bản đầu vào như là một chỉ số quan trọng. Phương pháp này khá phụ thuộc vào độ dài của văn bản đầu vào. Ví dụ, một từ xuất hiện ba lần trong một văn bản 10 từ có thể là từ quan trọng, nhưng có thể nó là một từ bình thường trong văn bản 1000 từ.

Xác suất của một từ  $w$ , ký hiệu là  $p(w)$ , được tính dựa trên số lần xuất hiện của từ  $w$ ,  $n(w)$ , trong toàn bộ các từ thuộc văn bản đầu vào  $N$ :

$$P(w) = n(w)/N$$

Hệ thống SumBasic [18] được phát triển dựa trên ý tưởng sử dụng xác suất của từ để tính toán câu quan trọng. Đối với mỗi câu  $\langle S_j \rangle$  trong văn bản đầu vào, nó gán một trọng số bằng xác suất trung bình của các từ chứa nội dung trong câu (một danh sách các từ không mang thông tin – stop words – sẽ bị loại khỏi quá trình đánh trọng số):

$$\text{Weight}(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{\| w_i \in S_j \|}$$

Tiếp theo, nó sẽ chọn các câu có điểm số tốt nhất, gồm những từ có xác suất cao nhất. Bước này đảm bảo rằng các từ có xác suất cao nhất đại diện cho chủ đề của văn bản đầu vào sẽ được đưa vào bản tóm tắt. Sau khi chọn một câu đưa vào tóm tắt, xác suất của mỗi từ trong câu được hiệu chỉnh:

$$p_{new}(w_i) = p_{old}(w_i)^2$$

Việc hiệu chỉnh này thể hiện rằng xác suất một từ xuất hiện hai lần trong bản tóm tắt là thấp hơn so với xác suất từ xuất hiện chỉ một lần. Quá trình lặp lại cho đến khi đạt được độ dài cần thiết của văn bản tóm tắt.

- Phương Pháp TF-IDF

Phương pháp dựa trên xác suất của từ phụ thuộc vào danh sách stop words để loại bỏ các từ không quan trọng khỏi bản tóm tắt. Việc quyết định từ nào sẽ đưa vào danh sách stop words sẽ ảnh hưởng tới hiệu năng của phương pháp word probability. Phương pháp TF-IDF (Term Frequency - Inverse Document Frequency) đã được nghiên cứu và phát triển để giải quyết hạn chế của phương pháp xác suất từ.

Phương pháp này sẽ đánh giá độ quan trọng của một từ bằng cách đánh trọng số cho từ. Các từ quan trọng trong văn bản sẽ được đánh trọng số cao, còn các từ phổ biến trong rất nhiều tài liệu (common words) sẽ được đánh trọng số thấp để loại bỏ khỏi danh sách đánh giá lựa chọn đưa vào văn bản tóm tắt. Trọng số của mỗi từ trong tài liệu d được tính như sau:

$$\text{Weight}(w) = fd(w) * \log \frac{D}{fD(w)}$$

Trong đó,  $fd(w)$  là term frequency của từ w trong tài liệu d,  $fD(w)$  là số tài liệu chứa từ w, và D là tổng số tài liệu. Như vậy, các từ xuất hiện trong hầu hết các tài liệu sẽ có giá trị IDF gần bằng 0. Trọng số TF IDF của từ là một chỉ số tốt để đánh giá mức độ quan trọng.

#### 2.4.2. Phương pháp đặc trưng đại diện

Phương pháp đặc trưng đại diện nhằm mô hình các đại diện của văn bản dựa trên một tập các đặc trưng và sử dụng chúng để xếp hạng các câu của văn bản đầu vào. Các phương pháp dựa trên đồ thị và kỹ thuật học máy thường được sử dụng để quyết định mức độ quan trọng của các câu sẽ đưa vào văn bản tóm tắt.

- Phương pháp đồ thị cho tóm tắt văn bản

Phương pháp dựa trên đồ thị thể hiện văn bản như là một đồ thị liên thông. Các câu tạo thành các đỉnh của đồ thị và các cạnh giữa các câu thể hiện sự liên quan giữa hai câu với nhau. Một kỹ thuật thường được sử dụng để nối hai đỉnh đó là đo lường sự tương đồng giữa hai câu và nếu nó lớn hơn một ngưỡng nhất định thì chúng liên thông nhau. Đồ thị này thể hiện kết quả ở hai phần: thứ nhất, một phần đồ thị con được tạo bao các chủ đề rời rạc trong văn bản; thứ hai, các câu được kết nối tới nhiều câu khác trong đồ thị là các câu quan trọng có thể lựa chọn đưa vào văn bản tóm tắt. Một phương pháp dựa trên đồ thị tiêu biểu đó là TextRank .

Phương pháp dựa trên đồ thị không cần các kỹ thuật xử lý ngôn ngữ tự nhiên đặc thù cho từng ngôn ngữ ngoài việc tách câu và từ, nên nó có thể áp dụng cho nhiều ngôn ngữ khác nhau.

- Kỹ thuật học máy cho tóm tắt văn bản

Phương pháp áp dụng học máy cho tóm tắt văn bản thực hiện giải bài toán phân loại nhị phân. Ý tưởng của chúng là phân loại các câu trong văn bản đầu vào thành hai tập là tập các câu tóm tắt và tập các câu không là tóm tắt dựa vào các đặc trưng mà chúng có. Tập dữ liệu huấn luyện gồm các văn bản và các bản tóm tắt trích chọn tương ứng.

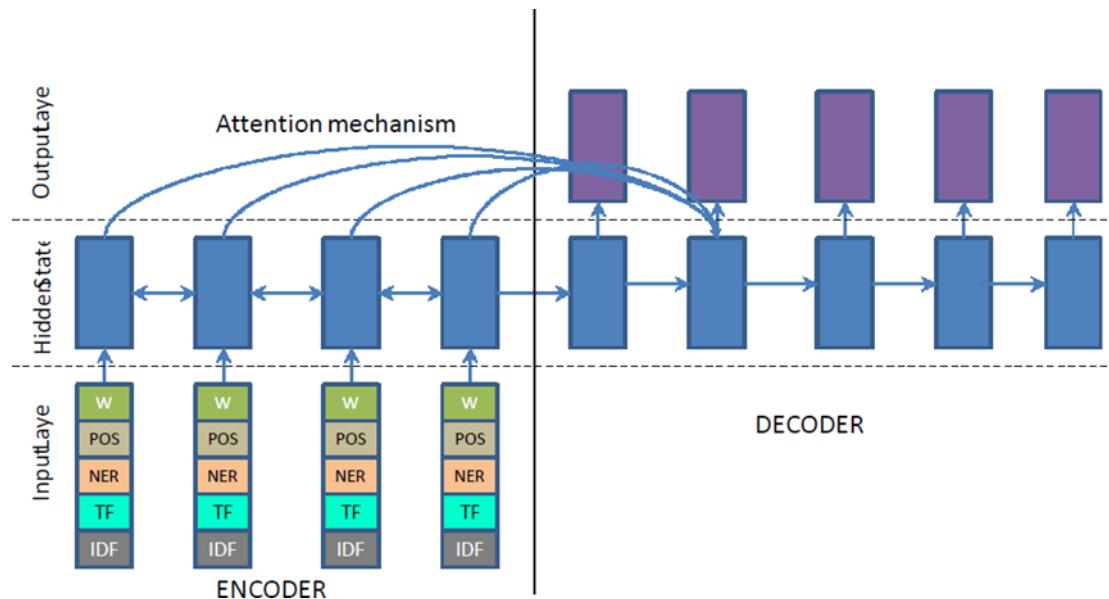
Xác suất một câu được chọn vào văn bản tóm tắt là điểm số của câu. Việc lựa chọn các hàm phân loại đóng vai trò quan trọng trong việc tính điểm cho các câu. Một số đặc trưng phân loại thường được sử dụng trong tóm tắt văn bản gồm có vị trí của câu trong văn bản, độ dài của câu, tồn tại của các từ viết hoa, độ tương đồng của câu với tiêu đề của văn bản... Có nhiều kỹ thuật học máy được áp dụng trong tóm tắt văn bản, tiêu biểu là áp dụng của mô hình Markov ẩn (Hidden Markov Model) [14].

## 2.5. Tóm tắt văn bản theo hướng tóm lược

Những năm gần đây với sự phát triển của phần cứng máy tính, cùng với nhiều kỹ thuật tiên tiến dựa trên mạng nơ-ron nhân tạo và kiến trúc mạng học sâu, một số nghiên cứu về tóm tắt văn bản bằng tóm lược đã được thực hiện với mục tiêu tạo được văn bản tóm tắt giống như cách con người thực hiện.

Nallapati và đồng nghiệp áp dụng mô hình chuỗi sang chuỗi (sequence-to-sequence) với cơ chế attention kết hợp với các đặc trưng ngôn ngữ (part-of-speech, named-entity và TF-IDF) để thực hiện tóm tắt văn bản theo hướng tóm lược (hình 2.1). Kết quả cho thấy mô hình có khả năng sinh ra các từ

không có trong văn bản đầu vào, nhiều ví dụ cho thấy mô hình có thể sinh ra đoạn tóm tắt gần giống với con người viết.



Hình 2.7 Mô hình sequence-to-sequence với cơ chế attention

Tác giả See và đồng nghiệp trong đề xuất cải tiến mô hình pointer-generator trên cơ sở mô hình chuỗi sang chuỗi, cho phép sao chép một (hoặc một số) từ văn bản nguồn vào văn bản tóm tắt trong trường hợp mô hình tạo ra một từ không xuất hiện trong tập từ vựng (unknown word). Mô hình này đã được thử nghiệm trên bộ dữ liệu tiếng Anh từ các bài báo của CNN/DailyMail, và cho thấy kết quả khá khả quan.

**Article:** smugglers lure arab and african migrants by offering discounts to get onto overcrowded ships if people bring more potential passengers, a cnn investigation has revealed.  
(...)

**Summary:** cnn investigation **uncovers** the **business inside** a **human smuggling ring**.

---

**Article:** eyewitness video showing white north charleston police officer michael slager shooting to death an unarmed black man has exposed discrepancies in the reports of the first officers on the scene. (...)

**Summary:** more **questions than answers emerge** in **controversial s.c.** police shooting.

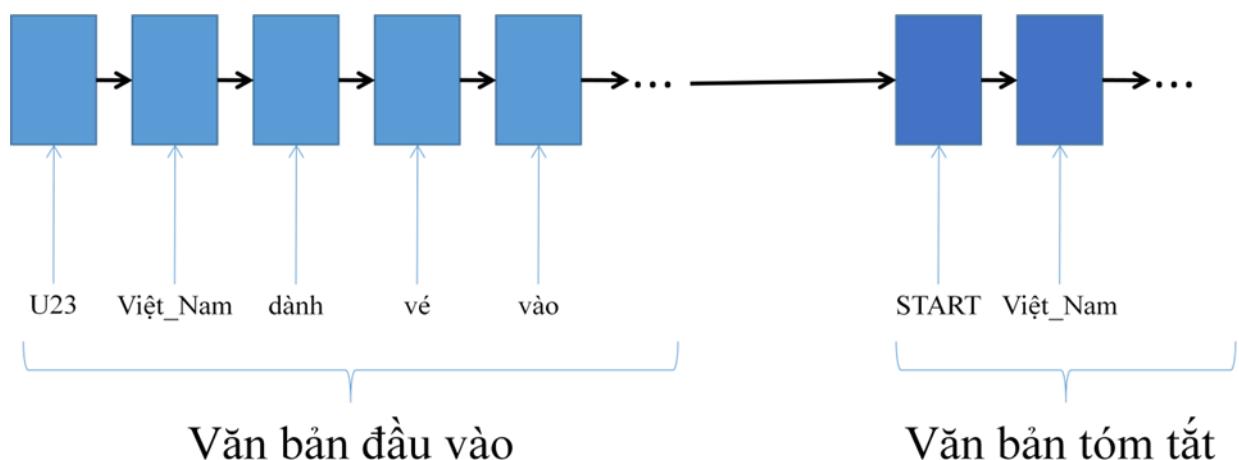
Hình 2.8 minh họa ví dụ về chạy thử nghiệm được tác giả công bố.

## CHƯƠNG 3: XÂY DỰNG ỦNG DỤNG VÀ KẾT QUẢ THỰC NGHIỆM

Bài toán tóm tắt văn bản theo hướng trích chọn được đặt ra như sau: đầu vào là một văn bản x gồm M từ: ( $x_1, x_2, x_3 \dots x_m$ ). Mục tiêu là tìm một chuỗi đầu ra y gồm N từ ( $N < M$ ) từ tập từ vựng V sao cho xác suất có điều kiện  $(P(y | x))$  là cực đại.

Quy trình này thường liên quan đến việc chọn lựa những đoạn văn bản quan trọng hoặc những câu chứa thông tin đặc biệt và tạo thành tóm tắt. Mỗi từ trong chuỗi đầu ra y có thể được chọn từ tập từ vựng V, và không nhất thiết phải xuất hiện trong chuỗi đầu vào x.

Cụ thể, để xây dựng hệ thống trích chọn, có thể sử dụng các mô hình máy học, đặc biệt là mô hình seq2seq (sequence-to-sequence) hoặc các kiến trúc chuyển đổi chú ý (attention-based architectures). Mô hình này sẽ học cách chọn lựa và sắp xếp các thành phần quan trọng của văn bản đầu vào để tạo ra một tóm tắt chính xác và ngắn gọn.



Hình 3.1. Mô hình bài toán tóm tắt văn bản

### 3.1. Các thư viện được sử dụng

#### 3.1.1. Thư viện Flask

Flask là một Web Framework rất nhẹ của Python, dễ dàng giúp người mới bắt đầu học Python có thể tạo ra website nhỏ. Flask cũng dễ mở rộng để xây dựng các ứng dụng web phức tạp.



Hình 3.2: Thư viện Flask

Flask có nền tảng là Werkzeug và Jinja2 và nó đã trở thành một trong những Web Framework phổ biến nhất của Python.

- Tính năng của Flask Framework:
  - Phát triển máy chủ
  - Phát triển trình gõ lỗi
  - Hỗ trợ sẵn sàng để kiểm thử đơn vị
  - Jinja2 templates
  - RESTful request dispatch
  - Hỗ trợ bảo mật cookie
  - Full WSGI compliant
  - Tài liệu mở rộng
  - Dựa trên Unicode
  - Khả năng tương thích công cụ dựa trên ứng dụng Google
  - Nhiều tiện ích mở rộng cho các tính năng mong muốn

- Tính modular và thiết kế gọn nhẹ
  - ORM-agnostic
  - Độ linh hoạt cao
  - Cung cấp xử lý HTTP request
  - API có đặc đáo và mạch lạc
  - Dễ dàng triển khai
- Ưu điểm của Flask Framework
    - Siêu nhỏ nhẹ, là một công cụ tối giản.
    - Tốc độ hoạt động cực nhanh.
    - Có khả năng hỗ trợ NoQuery.
    - Tương đối đơn giản (so với các framework có cùng chức năng khác).
    - Mang lại khả năng kết nối với các tiện ích mở rộng bởi không có ORM.
    - Trình duyệt được nhúng sẵn trình gõ rối.
    - Sử dụng các mã ngắn, đơn giản trong những bộ xương Python
    - Ngăn chặn các rủi ro về bảo mật khi lập trình web do ít phụ thuộc vào bên thứ ba.
    - Có khả năng kiểm soát mọi vấn đề khi dùng Flask.
    - Cho phép biên dịch mô-đun, thư viện, giúp việc lập trình nhanh chóng, dễ dàng hơn và không cần gõ code bậc thấp.
  - Nhược điểm của Flask Framework
 

Chính vì siêu nhỏ nhẹ và tối giản, Flask không phải là một lựa chọn tốt nếu lập trình viên muốn một framework có đầy đủ các tính năng. Thay vào đó, lập trình viên sẽ phải tự gọi các tiện ích mà mình có nhu cầu sử dụng vì nó không được tích hợp sẵn trong framework, và đôi khi việc này trở nên bất tiện và khiến cho khối lượng công việc phải làm tăng lên đáng kể.

### 3.1.2. Thư viện Numpy

NumPy là một thư viện lõi phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó. Numpy đặc biệt hữu ích khi thực hiện các hàm liên quan tới Đại Số Tuyến Tính. NumPy giải quyết vấn đề chậm một phần bằng cách cung cấp các mảng và hàm đa chiều và toán tử hoạt động hiệu quả trên mảng, sử dụng những điều này yêu cầu viết lại một số mã, chủ yếu là các vòng lặp bên trong, bằng cách sử dụng NumPy.



Hình 3.3: Thư viện Numpy

Dưới đây là một số đặc điểm và tính năng quan trọng của NumPy:

**Mảng đa chiều:** NumPy cho phép tạo và quản lý mảng đa chiều, tức là các mảng có số chiều lớn hơn 1. Điều này rất hữu ích cho việc làm việc với dữ liệu nhiều chiều như ma trận, tensor, hình ảnh và âm thanh.

**Hiệu suất cao:** NumPy được xây dựng dựa trên C/C++ nên có hiệu suất cao hơn so với các cấu trúc dữ liệu mảng trong Python chán. Nó cung cấp các phép toán vectorized, giúp thực hiện các phép toán trên toàn bộ mảng một cách nhanh chóng.

**Hỗ trợ các phép toán số học:** NumPy cung cấp một bộ các hàm số học như cộng, trừ, nhân, chia, mũ, căn bậc hai, logarit, và nhiều hàm toán học khác. Nhờ đó, việc thực hiện các phép toán số học trên mảng trở nên dễ dàng và hiệu quả.

Tích hợp với các thư viện khác: NumPy tương thích tốt với các thư viện phổ biến khác như SciPy (một thư viện mở rộng cho tính toán khoa học), Matplotlib (thư viện vẽ đồ thị) và Pandas (thư viện xử lý dữ liệu). Sự tích hợp này giúp xây dựng các ứng dụng phức tạp và thực hiện các phân tích dữ liệu một cách thuận tiện.

Thống kê và xử lý dữ liệu: NumPy cung cấp các hàm và phương thức mạnh mẽ để thực hiện các phép thống kê và xử lý dữ liệu trên mảng. Bạn có thể tính toán trung bình, phương sai, độ lệch chuẩn, tích vô hướng, tích ma trận, và nhiều tính toán thống kê khác trên mảng.

### 3.1.3. Thư viện werkzeug.utils



Hình 3.4: Thư viện Werkzeug

Thư viện werkzeug.utils là một phần của bộ công cụ Werkzeug, cung cấp các công cụ mạnh mẽ để phát triển ứng dụng web với Python. Hàm `secure_filename` được sử dụng để đảm bảo rằng các tệp tin được tải lên từ người dùng có tên an toàn, loại bỏ các ký tự không hợp lệ hoặc tiềm ẩn nguy cơ đối với hệ thống.

**Ứng dụng:** Khi người dùng tải tệp tin lên ứng dụng web, tệp tin có thể chứa các ký tự đặc biệt hoặc không an toàn. `secure_filename` đảm bảo rằng

tên tệp sẽ chỉ chứa các ký tự hợp lệ, giúp ngăn chặn các lỗ hổng bảo mật liên quan đến tên tệp và đường dẫn tệp.

#### 3.1.4. Thư viện tempfile



Hình 3.5: Thư viện tempfile

Tempfile là một thư viện tiêu chuẩn trong Python, được sử dụng để tạo ra các tệp tin và thư mục tạm thời một cách an toàn. Những tệp hoặc thư mục này sẽ chỉ tồn tại trong thời gian cần thiết và sẽ được tự động xóa sau khi không còn sử dụng.

**Ứng dụng:** Thư viện này hữu ích khi bạn cần làm việc với các tệp tin trong thời gian ngắn, chẳng hạn như lưu trữ tạm thời video, âm thanh, hoặc dữ liệu mà không cần lưu trữ lâu dài. tempfile đảm bảo rằng các tệp tạm thời được quản lý và xóa bỏ đúng cách, giúp tiết kiệm tài nguyên hệ thống và đảm bảo bảo mật dữ liệu.

#### 3.1.5. Thư viện underthesea



Hình 3.6: Thư viện Underthesea

Underthesea là một thư viện chuyên dụng cho xử lý ngôn ngữ tự nhiên (NLP) trong tiếng Việt. Thư viện này cung cấp các công cụ mạnh mẽ để phân tích, xử lý và khai thác ngôn ngữ tiếng Việt, bao gồm phân tách từ, phân tích cú pháp, gán nhãn từ loại, và nhiều tính năng khác.

Ứng dụng:

- Phân tách từ: Do tiếng Việt là ngôn ngữ không dấu câu rõ ràng giữa các từ, việc phân tách từ là một nhiệm vụ quan trọng trước khi xử lý ngôn ngữ.
- Gán nhãn từ loại: Xác định loại từ (danh từ, động từ, tính từ, v.v.) trong câu tiếng Việt để hỗ trợ phân tích cú pháp.
- Phân tích ngữ pháp: Xác định cấu trúc câu để hiểu rõ hơn về ý nghĩa của văn bản tiếng Việt.

### 3.1.6. Thư viện Numpy Networkx

NetworkX là một thư viện Python mạnh mẽ để tạo, thao tác và phân tích các cấu trúc đồ thị và mạng lưới. Nó đặc biệt hữu ích cho các bài toán liên quan đến lý thuyết đồ thị, mạng xã hội, hệ thống phân tán và các mạng lưới khác.

Ứng dụng:

- Phân tích mạng xã hội: NetworkX có thể được sử dụng để phân tích cấu trúc của các mạng xã hội, xác định những người có ảnh hưởng hoặc tính toán mức độ kết nối trong mạng.
- Mô hình hóa hệ thống: Thư viện giúp xây dựng các mô hình dựa trên lý thuyết đồ thị, ví dụ như mô hình giao thông, hệ thống phân phối, hay mạng máy tính.
- Thuật toán đồ thị: NetworkX cung cấp các thuật toán như tìm đường ngắn nhất, phát hiện chu trình trong đồ thị, và phân tích độ trung tâm của các nút trong mạng.

### 3.1.7. Thư viện moviepy.editor



Hình 3.7: Thư viện MoviePy

MoviePy là một thư viện đa phương tiện mạnh mẽ trong Python, hỗ trợ xử lý và chỉnh sửa video. MoviePy giúp bạn dễ dàng thao tác với video như cắt, ghép, thêm hiệu ứng, hay chuyển đổi giữa các định dạng khác nhau.

Ứng dụng:

- Cắt video: Bạn có thể sử dụng MoviePy để cắt bỏ các đoạn không mong muốn trong video.
- Thêm âm thanh: MoviePy hỗ trợ thêm, tách, và chỉnh sửa các đoạn âm thanh trong video.
- Chuyển đổi định dạng video: MoviePy có thể chuyển đổi video giữa nhiều định dạng khác nhau như MP4, AVI, GIF.
- Làm việc với khung hình: Thư viện cho phép thao tác từng khung hình của video để tạo ra các hiệu ứng độc đáo.

### 3.1.8. Thư viện SpeechRecognition



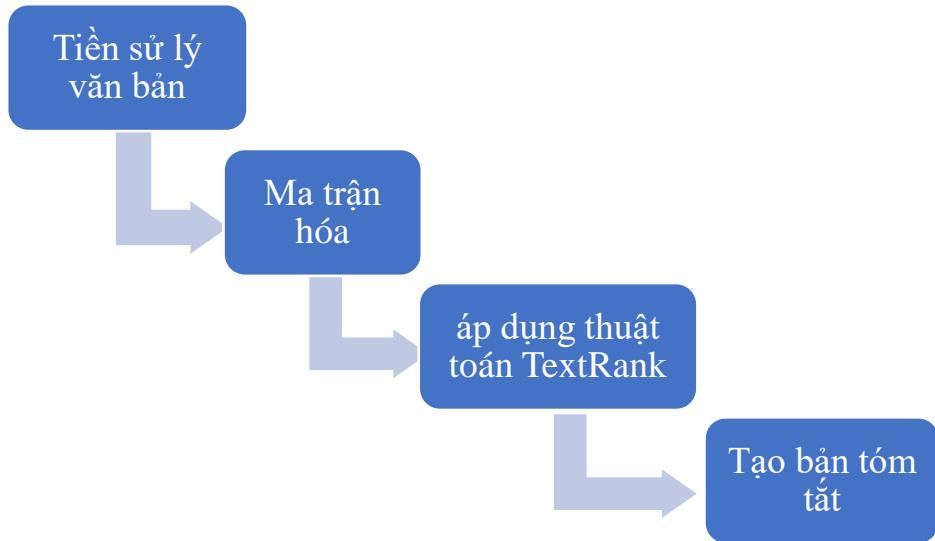
Hình 3.8: Thư viện SpeechRecognition

SpeechRecognition là một thư viện Python mạnh mẽ dùng để nhận dạng giọng nói, chuyển đổi âm thanh từ giọng nói thành văn bản. Thư viện này hỗ trợ nhiều dịch vụ nhận dạng giọng nói khác nhau như Google Speech API, IBM, Microsoft, và Sphinx.

#### Ứng dụng:

- Nhận dạng giọng nói: Thư viện này có thể chuyển đổi trực tiếp âm thanh từ microphone hoặc từ tệp âm thanh thành văn bản.
- Tích hợp với API: SpeechRecognition có thể kết hợp với các API như Google Speech để nhận dạng giọng nói trực tuyến với độ chính xác cao.
- Xử lý âm thanh: Bạn có thể sử dụng thư viện này để phân tích các tệp âm thanh và nhận dạng ngôn ngữ nói trong các ứng dụng như trợ lý ảo, hệ thống điều khiển bằng giọng nói.

### 3.2. Tổng quan ứng dụng tóm tắt văn bản



- Giới Thiệu

Trong thời đại thông tin, việc xử lý và tổng hợp thông tin nhanh chóng và chính xác trở nên cực kỳ quan trọng. Ứng dụng tóm tắt văn bản mà em phát triển nhằm mục đích giải quyết vấn đề này bằng cách cung cấp bản tóm tắt ngắn gọn, súc tích từ văn bản gốc. Qua đó, giúp người đọc nắm bắt nhanh chóng thông tin cần thiết mà không cần phải đọc toàn bộ văn bản.

- Tokenization và Tiền Xử Lý

Bước đầu tiên trong quy trình xử lý văn bản của em là "Tokenization", sử dụng word\_tokenize từ thư viện underthesea. Quy trình này phân tách văn bản thành các đơn vị nhỏ hơn như câu và từ. Tiếp theo, em thực hiện tiền xử lý bằng cách loại bỏ các ký tự không cần thiết và chuẩn hóa ngôn ngữ, bao gồm việc loại bỏ các từ dừng, nhằm mục đích tinh giản văn bản và chuẩn bị cho các bước phân tích tiếp theo.

- Ma trận hóa

Sau khi hoàn tất quá trình tokenization và tiền xử lý, em sử dụng phương pháp TextRank để biểu diễn các câu dựa trên độ tương đồng về từ vựng. Cụ thể, em tính toán mức độ tương đồng giữa mỗi cặp câu dựa trên số lượng từ chung giữa chúng. Kết quả này được thể hiện dưới dạng một ma trận tương đồng, trong đó mỗi giá trị biểu thị mức độ giống nhau giữa hai câu. Những câu có nhiều từ chung sẽ có độ tương đồng cao hơn, điều này giúp em hiểu cách các câu liên kết với nhau về mặt nội dung..

- Tính Toán Độ Tương Đồng Bằng Cosine Similarity

Để xác định mức độ tương đồng giữa các câu, em sử dụng một công thức dựa trên số lượng từ chung giữa chúng. Cosine similarity không còn được áp dụng trong mô hình TextRank đơn giản này. Thay vào đó, công thức tính dựa trên tỷ lệ của từ chung với độ dài của các câu. Ma trận tương đồng này là nền tảng để tạo ra một đồ thị liên kết, trong đó các câu là các nút và độ tương đồng giữa chúng tạo nên các cạnh liên kết. Sau đó, thuật toán PageRank sẽ được áp dụng lên đồ thị này để xác định mức độ quan trọng của từng câu, hỗ trợ trong việc tóm tắt văn bản bằng cách chọn ra những câu có giá trị cao nhất.

- Áp Dụng TextRank Algorithm

TextRank, một thuật toán dựa trên cơ sở của PageRank, được sử dụng để xác định những câu quan trọng trong văn bản. Thuật toán này phân tích mối quan hệ giữa các câu và xếp hạng chúng dựa trên mức độ quan trọng. Những câu có xếp hạng cao nhất, tức là có nhiều liên kết với các câu khác và chứa thông tin trung tâm, được chọn để tạo nên đoạn tóm tắt.

- Quy Trình Tạo Ra Tóm Tắt

Kết hợp những câu được chọn từ bước trước, em tạo ra đoạn văn tóm tắt. Quy trình này không chỉ đơn giản là lựa chọn câu, mà còn bao gồm việc

sắp xếp lại chúng một cách logic và mạch lạc để tạo ra một đoạn văn có ý nghĩa và dễ hiểu. Chất lượng của tóm tắt được đánh giá dựa trên mức độ giữ lại thông tin quan trọng và sự mạch lạc của văn bản.

- **Kết Luận và Hướng Phát Triển**

Kết luận của báo cáo này nhấn mạnh vào việc ứng dụng tóm tắt văn bản đã đạt được mục tiêu cung cấp bản tóm tắt ngắn gọn, chính xác từ văn bản gốc. em cũng đề cập đến những hạn chế hiện tại và hướng phát triển trong tương lai, bao gồm việc cải thiện chất lượng tóm tắt và khả năng xử lý ngôn ngữ tự nhiên phức tạp hơn.

### 3.3. Cài đặt ứng dụng tóm tắt văn bản

#### 3.3.1. Mô hình giải quyết bài toán

Trong phần này, em sẽ xem xét chi tiết cách thức mà ứng dụng tóm tắt văn bản được cài đặt để giải quyết bài toán tóm tắt văn bản. Mô hình giải quyết bài toán bao gồm các bước từ tiền xử lý dữ liệu đến tạo ra tóm tắt cuối cùng.

- **Tiền Xử Lý Dữ liệu**

Bước đầu tiên trong quá trình tóm tắt văn bản là tiền xử lý. Điều này bao gồm việc tách văn bản thành từng câu và từng từ, loại bỏ các ký tự không cần thiết, chuẩn hóa ngôn ngữ, và loại bỏ các từ dừng. Tiền xử lý giúp loại bỏ thông tin không cần thiết và chuẩn bị dữ liệu cho quá trình phân tích.

- **Biểu Diễn Văn Bản**

Sau khi hoàn tất tiền xử lý, từng từ trong văn bản được chuyển đổi thành ma trận sử dụng mô hình TextRank. Điều này cho phép em biểu diễn mỗi câu được đánh giá dựa trên mối liên hệ và mức độ tương đồng với các câu khác trong văn bản, tạo điều kiện cho việc xử lý bằng các thuật toán máy học.

### - Xác Định Độ Tương Đồng

Sử dụng phương pháp cosine similarity, em đánh giá mức độ tương đồng giữa các câu trong văn bản. Điều này giúp xác định những câu có nội dung tương tự nhau và loại bỏ thông tin trùng lặp trong quá trình tạo tóm tắt.

### - Xếp Hạng và Lựa Chọn Câu

Sử dụng thuật toán TextRank, em xác định và xếp hạng các câu quan trọng dựa trên mức độ liên kết và tương đồng với các câu khác trong văn bản. Các câu có xếp hạng cao nhất được chọn làm phần cốt lõi của bản tóm tắt.

### - Tạo Ra Tóm Tắt Cuối Cùng

Cuối cùng, các câu được chọn từ bước trước được tổng hợp lại để tạo ra bản tóm tắt. Quá trình này đòi hỏi việc sắp xếp lại các câu một cách logic và mạch lạc để đảm bảo rằng tóm tắt không chỉ ngắn gọn mà còn dễ hiểu và chính xác về mặt thông tin.

Tổng Kết Mô hình giải quyết bài toán của em kết hợp cả kỹ thuật tiền xử lý dữ liệu, biểu diễn văn bản dưới dạng vector, phân tích độ tương đồng, xếp hạng câu, và cuối cùng là tạo ra tóm tắt. Sự kết hợp này tạo ra một cách tiếp cận toàn diện trong việc tóm tắt văn bản, giúp người dùng dễ dàng nắm bắt thông tin quan trọng mà không cần phải đọc toàn bộ văn bản.

#### 3.3.2. Tiền xử lý văn bản

Tiền xử lý văn bản là một bước quan trọng trong quy trình tóm tắt văn bản, nhằm chuẩn bị dữ liệu cho các quá trình xử lý sau. Quy trình này bắt đầu với tokenization, sử dụng thư viện `underthesea`, để chia văn bản thành các đơn vị nhỏ như câu và từ. Điều này giúp phân tách và xác định cấu trúc cơ bản của văn bản, cung cấp nền tảng cho việc phân tích ngữ nghĩa. Tiếp theo, em loại bỏ các ký tự không cần thiết như dấu chấm câu và số, đồng thời chuẩn

hóa văn bản bằng cách chuyển tất cả chữ cái thành chữ thường để tránh sự không nhất quán. Một phần quan trọng khác của quá trình này là loại bỏ các từ dừng - những từ thường xuất hiện trong ngôn ngữ nhưng không mang nhiều ý nghĩa ngữ nghĩa, như 'và', 'là', 'của'. Bằng cách loại bỏ chúng, em tập trung vào các từ quan trọng mang thông tin chính của văn bản. Quy trình này giúp tinh giản văn bản, loại bỏ thông tin dư thừa, và chuẩn bị dữ liệu một cách hiệu quả cho các bước phân tích và tóm tắt văn bản tiếp theo.

### 3.3.3. Xây dựng mô hình TextRank

Trong quá trình tóm tắt văn bản, việc hiểu và xử lý sự liên kết giữa các câu là một phần không thể thiếu, và mô hình TextRank đóng vai trò quan trọng trong việc này. TextRank là một thuật toán dựa trên đồ thị, được sử dụng để xác định mức độ quan trọng của các câu trong văn bản thông qua việc đánh giá sự tương đồng giữa chúng.

Cụ thể, trong dự án của chúng tôi, mô hình TextRank được xây dựng dựa trên việc tính toán mức độ tương đồng giữa các câu trong văn bản gốc. Quá trình này bắt đầu bằng cách tạo ra một ma trận tương đồng, trong đó mỗi cặp câu sẽ được so sánh với nhau dựa trên số lượng từ vựng chung. Mỗi câu sẽ được xem như một "nút" trong đồ thị, và mỗi liên kết giữa các câu được thể hiện qua độ tương đồng của chúng, tạo thành các cạnh nối giữa các nút.

Các bước chính trong quá trình triển khai TextRank bao gồm:

- Xây dựng ma trận tương đồng: Sau khi các câu đã được tiền xử lý, em tính toán mức độ tương đồng giữa các câu bằng cách sử dụng số lượng từ chung giữa chúng. Ma trận tương đồng này sẽ làm đầu vào để xây dựng đồ thị liên kết giữa các câu.
- Tạo đồ thị liên kết: Mỗi câu được biểu diễn dưới dạng một nút trong đồ thị, và các cạnh giữa các nút được xác định dựa trên giá trị trong ma

trận tương đồng. Các câu có mức độ tương đồng cao sẽ có liên kết mạnh hơn trong đồ thị.

- Áp dụng thuật toán TextRank: Thuật toán PageRank được áp dụng để xếp hạng các câu dựa trên mức độ liên kết và quan trọng của chúng trong toàn bộ văn bản. Câu nào có mức độ liên kết mạnh với nhiều câu khác sẽ được đánh giá là quan trọng hơn.
- Lựa chọn câu quan trọng: Sau khi xếp hạng các câu, những câu có điểm số cao nhất sẽ được lựa chọn làm phần cốt lõi của bản tóm tắt. Những câu này thường chứa thông tin quan trọng và có tính đại diện cao cho nội dung của văn bản gốc.

Mô hình TextRank, với khả năng phân tích và đánh giá mức độ quan trọng của các câu trong văn bản, trở thành một công cụ hữu ích trong việc tạo ra bản tóm tắt chất lượng cao. Thay vì chỉ biểu diễn ngữ nghĩa từng từ như Word2Vec, TextRank tập trung vào sự liên kết giữa các câu, giúp xác định thông tin quan trọng một cách chính xác và hiệu quả hơn trong bối cảnh của toàn bộ văn bản.

### 3.3.4. Xây dựng ma trận tương đồng

Trong quá trình tóm tắt văn bản, việc đánh giá mức độ tương đồng giữa các câu là một bước quan trọng, và để thực hiện điều này, em sử dụng phương pháp cosine similarity. Cosine similarity là một kỹ thuật phổ biến trong xử lý ngôn ngữ tự nhiên, cho phép em đo lường mức độ tương đồng ngữ nghĩa giữa các câu dựa trên các vector đại diện của chúng.

Quy trình xây dựng ma trận tương đồng diễn ra như sau:

- Xây dựng ma trận tương đồng: Sau khi các câu đã được tiền xử lý, em tính toán mức độ tương đồng giữa các câu bằng cách sử dụng số lượng từ chung

giữa chúng. Ma trận tương đồng này sẽ làm đầu vào để xây dựng đồ thị liên kết giữa các câu.

- Tạo đồ thị liên kết: Mỗi câu được biểu diễn dưới dạng một nút trong đồ thị, và các cạnh giữa các nút được xác định dựa trên giá trị trong ma trận tương đồng. Các câu có mức độ tương đồng cao sẽ có liên kết mạnh hơn trong đồ thị.
- Áp dụng thuật toán TextRank: Thuật toán PageRank được áp dụng để xếp hạng các câu dựa trên mức độ liên kết và quan trọng của chúng trong toàn bộ văn bản. Câu nào có mức độ liên kết mạnh với nhiều câu khác sẽ được đánh giá là quan trọng hơn.
- Lựa chọn câu quan trọng: Sau khi xếp hạng các câu, những câu có điểm số cao nhất sẽ được lựa chọn làm phần cốt lõi của bản tóm tắt. Những câu này thường chứa thông tin quan trọng và có tính đại diện cao cho nội dung của văn bản gốc.

Mô hình TextRank, với khả năng phân tích và đánh giá mức độ quan trọng của các câu trong văn bản, trở thành một công cụ hữu ích trong việc tạo ra bản tóm tắt chất lượng cao. Thay vì chỉ biểu diễn ngữ nghĩa từng từ như Word2Vec, TextRank tập trung vào sự liên kết giữa các câu, giúp xác định thông tin quan trọng một cách chính xác và hiệu quả hơn trong bối cảnh của toàn bộ văn bản.

Biểu diễn mối liên kết giữa các câu:

- Dựa trên phương pháp TextRank, thay vì biểu diễn từng câu dưới dạng vector như Word2Vec, em phân tích mối liên kết giữa các câu dựa trên số lượng từ vựng chung giữa chúng. Mỗi câu được xem như một nút trong đồ thị, và các cạnh liên kết giữa các câu được xác định thông qua

việc tính toán số lượng từ chung, cho thấy sự liên kết nội dung giữa các câu.

Tính toán mức độ tương đồng giữa các câu:

- Với mỗi cặp câu trong văn bản, em tính toán mức độ tương đồng bằng cách đếm số lượng từ chung giữa hai câu. Sự tương đồng này phản ánh mức độ nội dung mà hai câu chia sẻ, giúp xác định các câu có ngữ cảnh hoặc thông tin tương tự nhau. Điều này tương tự như việc tính toán cosine similarity, nhưng thay vì dùng vector, em dựa trên sự trùng lặp từ ngữ.

Xây dựng ma trận tương đồng:

- Kết quả của các phép tính trên được sử dụng để xây dựng một ma trận tương đồng (similarity\_matrix), trong đó mỗi phần tử `similarity_matrix[i][j]` biểu diễn mức độ tương đồng giữa câu thứ i và câu thứ j trong văn bản. Ma trận này là vuông, với kích thước bằng số lượng câu trong văn bản. Các phần tử trên đường chéo chính thường được đặt bằng 0 vì nó biểu diễn sự tương đồng của một câu với chính nó.
- Ma trận tương đồng này cung cấp một cái nhìn toàn diện về mức độ liên kết giữa các câu trong văn bản, giúp em xác định những câu nào có vai trò quan trọng và cần được đưa vào bản tóm tắt cuối cùng. Phương pháp này đặc biệt hữu ích trong việc xác định các câu mang thông tin tương tự nhau, từ đó giúp loại bỏ thông tin trùng lặp và tập trung vào những nội dung cốt lõi của văn bản.

### 3.3.5. Áp dụng thuật toán textrank và tạo tóm tắt

Ở bước cuối cùng trong quá trình tạo tóm tắt văn bản, em áp dụng thuật toán TextRank, một phương pháp được phát triển dựa trên nguyên lý của thuật toán PageRank của Google. TextRank được thiết kế để xác định những câu quan trọng nhất trong văn bản dựa trên mối quan hệ tương đồng giữa chúng.

TextRank hoạt động theo cách sau:

1. Xây Dựng Đồ Thị: Mỗi câu trong văn bản được biểu diễn như một nút trên đồ thị. Các cạnh nối giữa các nút đại diện cho mức độ tương đồng giữa các câu, được định lượng từ ma trận tương đồng đã xây dựng trước đó.
2. Tính Điểm TextRank: Điểm TextRank của mỗi câu được tính toán dựa trên tổng điểm tương đồng của nó với các câu khác. Điều này được thực hiện bằng cách sử dụng công thức  $scores = np.sum(similarity\_matrix, axis=1)$ . Mỗi câu nhận được một điểm số dựa trên mức độ "quan trọng" của nó trong văn bản, tức là mức độ liên kết của nó với các câu khác.
3. Xếp Hạng và Lựa Chọn Câu: Dựa trên điểm số này, các câu được xếp hạng từ cao xuống thấp. Những câu có điểm số cao nhất được xem là những câu mang nhiều thông tin quan trọng nhất và sẽ được lựa chọn để tạo thành bản tóm tắt.
4. Tạo Tóm Tắt Cuối Cùng: Cuối cùng, những câu được chọn này được kết hợp lại một cách logic và mạch lạc để tạo ra đoạn văn tóm tắt. Quá trình này không chỉ dựa trên điểm số, mà còn cân nhắc đến trật tự logic và mối quan hệ ngữ nghĩa giữa các câu, đảm bảo rằng tóm tắt cuối cùng không chỉ ngắn gọn mà còn dễ hiểu và chính xác về mặt thông tin.

TextRank là một công cụ mạnh mẽ, cho phép tự động xác định và trích xuất những phần quan trọng nhất của văn bản, giúp tóm tắt văn bản trở nên chính xác và hiệu quả hơn. Sự kết hợp của TextRank với các phương pháp xử lý ngôn ngữ tự nhiên khác trong ứng dụng của em giúp tạo ra bản tóm tắt có chất lượng cao, phản ánh chính xác nội dung và thông điệp chính của văn bản gốc.

### 3.4. Kết quả thực nghiệm

#### 3.4.1. Môi trường thực nghiệm:

- Công cụ xây dựng ứng dụng demo:

Em đã phát triển và thử nghiệm ứng dụng tóm tắt văn bản trên một máy tính có cấu hình như sau:

- CPU: AMD ryzen 5 4600u , 12CPUs ~2.1GHz
- RAM: 8GB.
- GPU: AMD Radeon Graphics
- Hệ Điều Hành: Windows 11 Pro.
- Ngôn Ngữ Lập Trình: Python, sử dụng trình biên dịch Python

Các công cụ chính được sử dụng trong việc xây dựng ứng dụng bao gồm:

- PyCharm Community 2023: Một môi trường phát triển tích hợp (IDE) hiệu quả cho Python.

Các bước xây dựng ứng dụng tóm tắt văn bản:

#### 1. Chuẩn Bị Mô Hình và Dữ Liệu:

- Tạo file summarizer.py:
- Định nghĩa hàm build\_similarity\_matrix: Đây là hàm để tính toán ma trận tương đồng giữa các câu trong văn bản. Em cần sử dụng kỹ thuật đếm số lượng từ chung giữa các câu để xác định mức độ tương đồng giữa chúng. Ma trận này sẽ được sử dụng làm đầu vào cho thuật toán TextRank
- Định nghĩa hàm summarize\_text để tóm tắt văn bản dựa trên các kỹ thuật đã mô tả trước đó.

#### 2. Xây dựng ứng dụng flask:

- Tạo file app.py, models.py, users.py:
- Import các thư viện và modules cần thiết (Flask, gensim, underthesea, numpy, ...).
- Thiết lập Flask app và cấu hình đường dẫn.
- Tạo route / để hiển thị trang chủ với form nhập văn bản.
- Tạo route /summarize để xử lý tóm tắt văn bản được nhập.
- Chạy ứng dụng Flask.

### 3. Thiết kế trang html:

- Tạo file index.html, login.html, register.html:

Thiết kế giao diện cho trang chủ, đăng nhập, đăng ký và bao gồm form nhập văn bản và nút gửi.

Hiển thị kết quả tóm tắt sau khi người dùng nhập văn bản.

### 4. CSS và JavaScript:

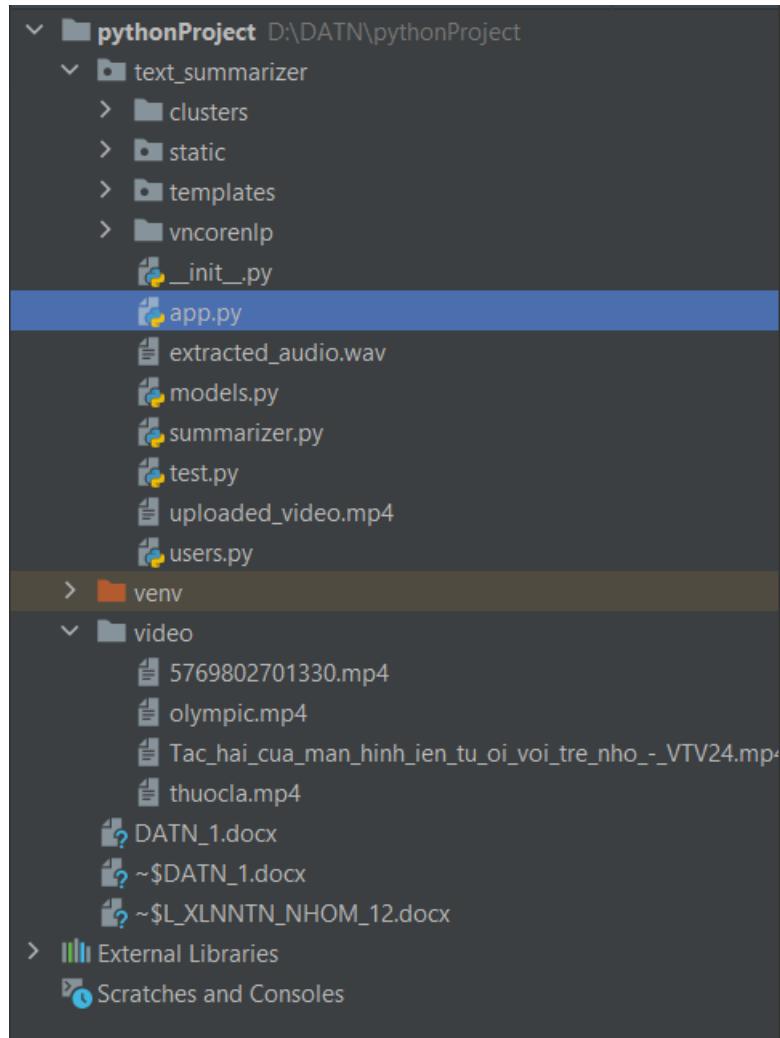
- Tạo file static/styles.css, login.css, register.css:

Thiết kế giao diện bằng CSS, tùy chỉnh giao diện theo ý muốn.

### 5. Chạy ứng dụng:

- Chạy Flask app bằng lệnh python app.py.
- Truy cập trình duyệt và vào địa chỉ http://127.0.0.1:5000/ để sử dụng ứng dụng.

Cấu trúc thư mục:



Hình 3.9 Cấu trúc ứng dụng tóm tắt văn bản

#### 3.4.2. Quá trình thử nghiệm:

- Load dữ liệu: Sử dụng bộ dữ liệu văn bản mẫu để xây dựng mô hình tóm tắt văn bản dựa trên thuật toán TextRank.
- Chuẩn bị đặc trưng: Thực hiện tokenization và tiền xử lý văn bản, bao gồm việc tách câu và từ trong văn bản. Đây là bước chuẩn bị cho việc xây dựng ma trận tương đồng giữa các câu.
- Xây dựng ma trận tương đồng: Sử dụng các phương pháp đã mô tả trước đó để tính toán độ tương đồng giữa các câu dựa trên số lượng từ chung. Ma trận

này biểu diễn mức độ liên kết giữa các câu, giúp xác định câu nào quan trọng trong văn bản.

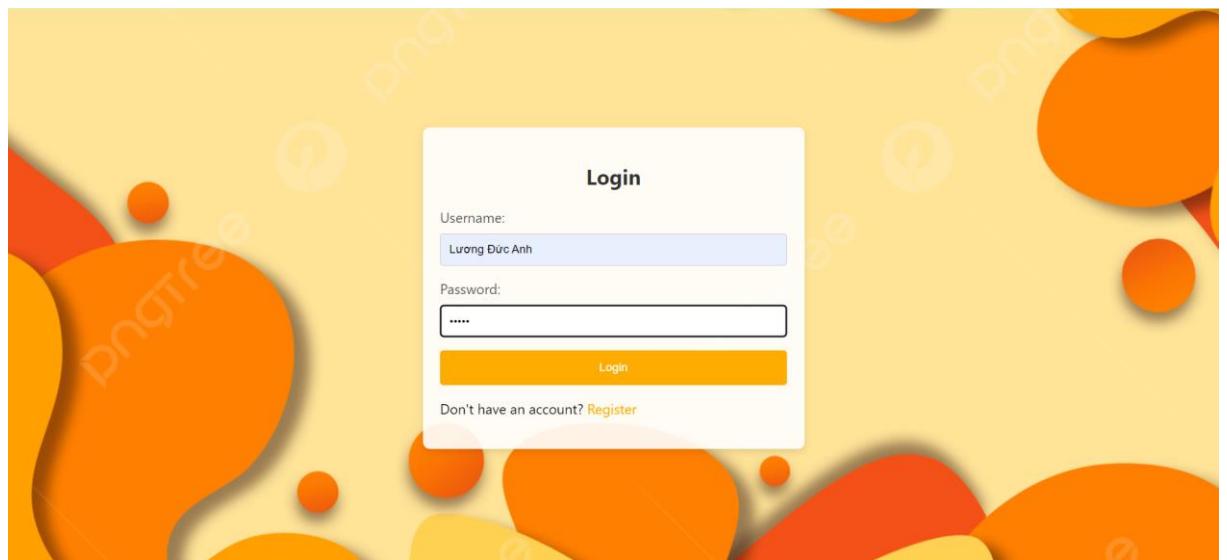
- Áp dụng thuật toán TextRank: Sử dụng thuật toán TextRank để xếp hạng các câu dựa trên độ tương đồng trong ma trận. Những câu có điểm số cao nhất sẽ được chọn làm phần cốt lõi của bản tóm tắt.
- Dự đoán và đánh giá: Sử dụng mô hình TextRank để tạo tóm tắt cho tập dữ liệu văn bản kiểm thử.

Đánh giá chất lượng tóm tắt dựa trên các tiêu chí như độ chính xác, tính mạch lạc và tính ngắn gọn của bản tóm tắt so với văn bản gốc.

#### 3.4.3. Dự đoán và hiển thị kết quả:

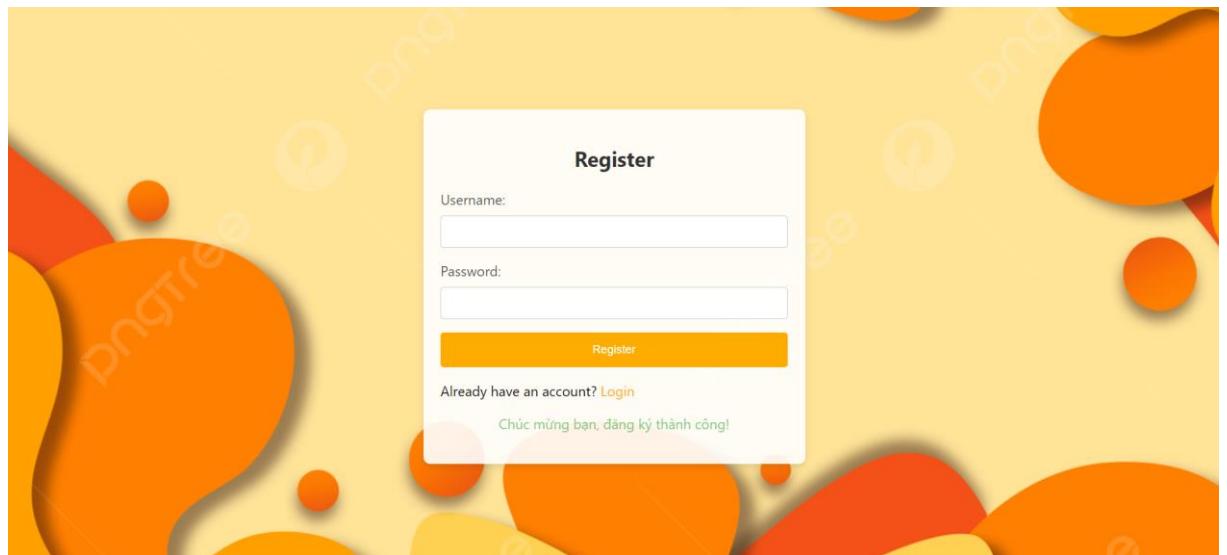
*Giao diện ứng dụng:*

+ Giao diện Login:



Hình 3.10 Giao diện Login

+Giao diện Register:



Hình 3.11 Giao diện Login

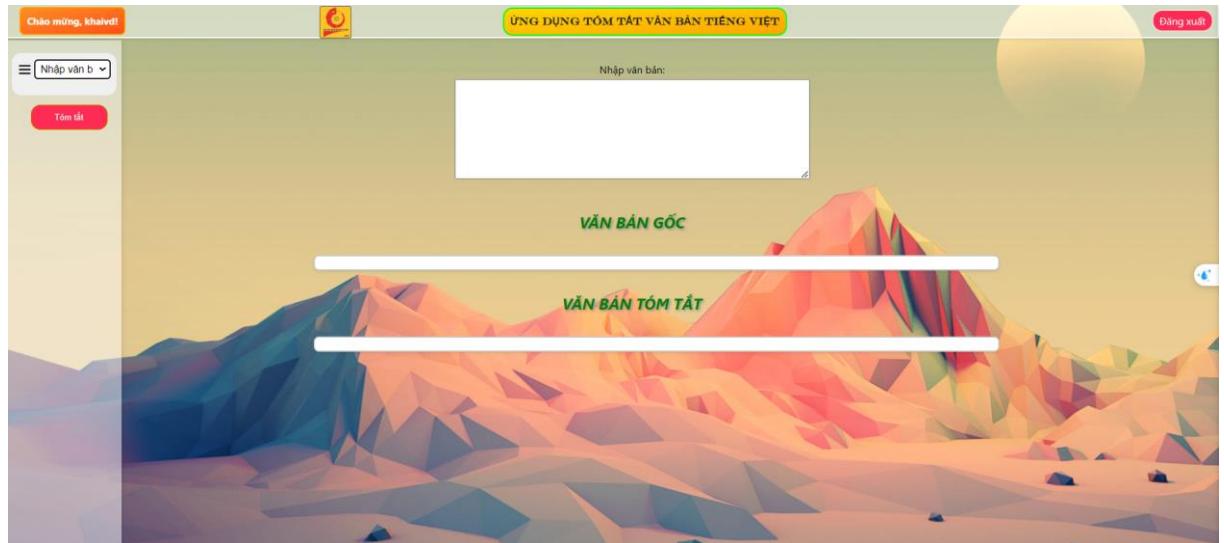
+Giao diện chính:



Hình 3.12 Giao diện trang chủ

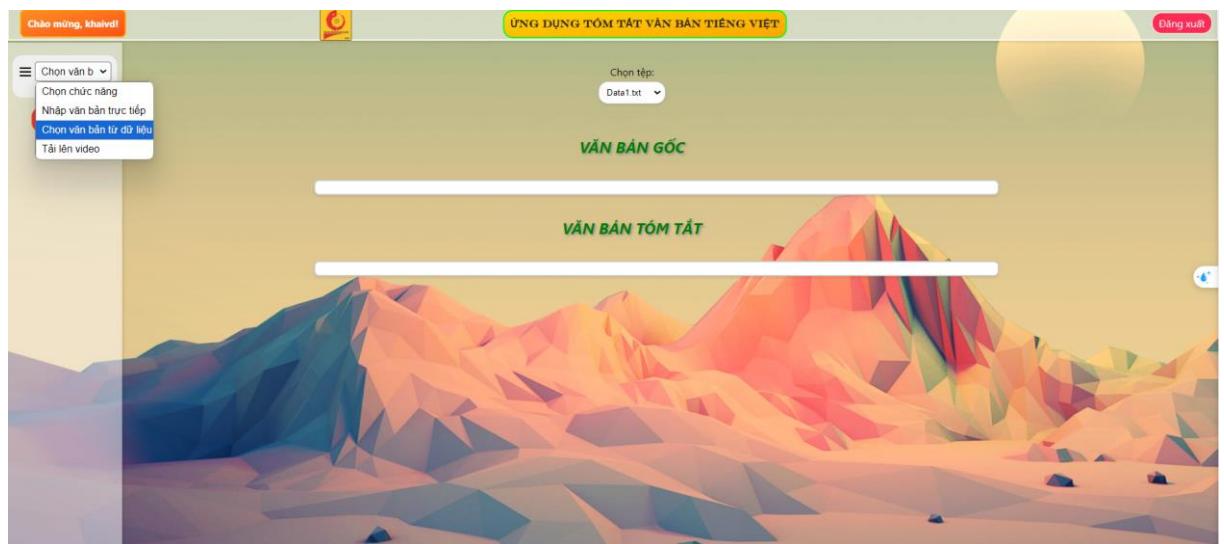
### Các chức năng chính:

TH1 Nhập trực tiếp văn bản để test



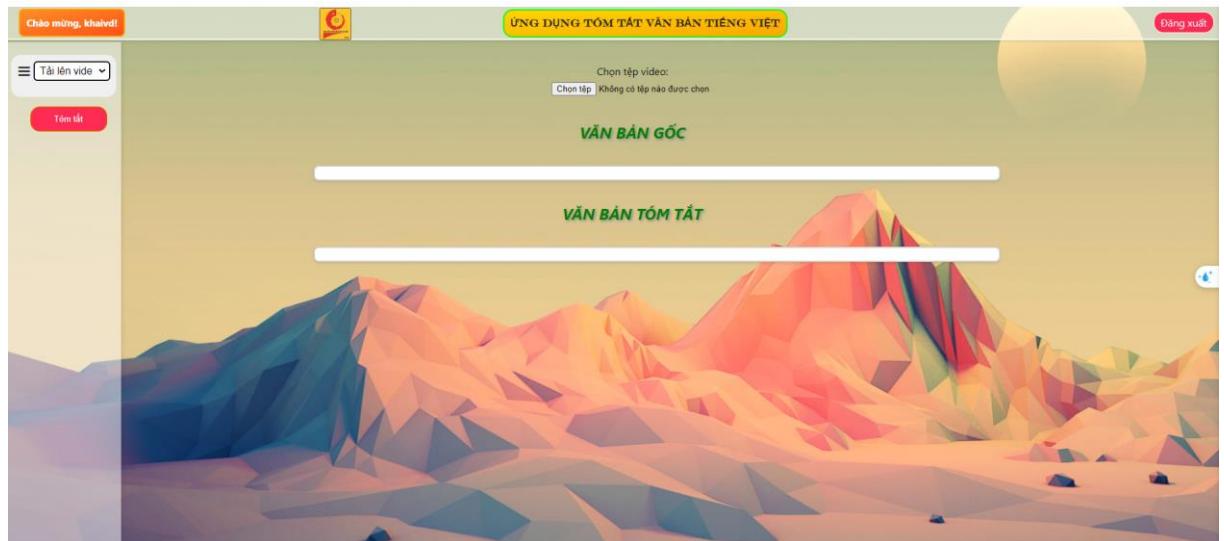
Hình 3.13 TH1 Nhập trực tiếp văn bản để test

TH2 Lấy văn bản từ Data để test



Hình 3.14 TH2 Lấy văn bản từ Data để test

### TH3 Lấy văn bản từ video để test



Hình 3.15 TH3 lấy văn bản từ video để test

### Kết quả TH1 nhập trực tiếp văn bản để test

Hình 3.16 Test th1

### Kết quả TH2 Lấy văn bản từ Data để test

## VĂN BẢN GỐC

Ngày 23/10, Trung\_Quốc và Ấn Độ đã ký một thỏa\_thuận nhằm làm giảm căng\_thẳng ở khu\_vực biên\_giới . Thỏa\_thuận đạt được là kết\_quả của những nỗ\_lực không ngừng giữa 2 cường\_quốc nhằm phá vỡ bế\_tắc vốn tồn\_tại suốt nhiều thập\_kỷ qua liên\_quan đến vùng lãnh\_thổ mà\_cả 2 cùng tuyên\_bố chủ\_quyền trên dãy Himalaya . Thủ\_tướng Ấn Độ\_Manmohan\_Singh và Thủ\_tướng Trung\_Quốc\_Lý\_Khắc\_Cường đã có cuộc hội\_dàm ở Bắc\_Kinh ngày 23/10 (Ảnh : Reuters ) Thỏa\_thuận trên được ký\_kết tại Đại\_Lễ\_dường Nhàn\_dân ở Bắc\_Kinh sau cuộc họp giữa Thủ\_tướng Ấn Độ\_Manmohan\_Singh và Thủ\_tướng Trung\_Quốc\_Lý\_Khắc\_Cường . Trung\_Quốc tuyên\_bố chủ\_quyền với hơn 90.000km<sup>2</sup> ở khu\_vực 2 nước đang tranh\_chấp ở phía đông của dãy Himalaya . Trong khi đó , Ấn Độ cho rằng , Trung\_Quốc chiếm 38.000 km<sup>2</sup> lãnh\_thổ của nước này trên cao\_nguyên Aksai\_Chin ở phía tây Himalaya . Trong lịch\_sử , 2 nước đã có một cuộc chiến\_tranh biên\_giới ngắn ngày vào năm 1962 khiến mối quan\_hệ sau đó giữa Trung\_Quốc và Ấn Độ luôn ở trong trạng\_thái ngõ\_vực . Đầu năm nay , Ấn Độ đã cáo\_buộc quân\_đội Trung\_Quốc xâm\_phạm lãnh\_thổ của nước này dưới danh\_nghĩa tiến\_hành các hoạt động tuẫn\_trá . Trả\_lời các phóng\_viên sau cuộc hội\_dàm với Thủ\_tướng Singh , ông Lý\_Khắc\_Cường nói : " Tôi chắc\_chắn thỏa\_thuận biên\_giới này sẽ giúp duy\_trì hòa\_bình , ổn định trong khu\_vực biên\_giới của chúng\_tôi ". Một quan\_chức Ấn Độ tuẫn trước cho\_biết , thỏa\_thuận hợp\_tác quốc\_phóng biên\_giới được xây\_dựng dựa trên các biện\_pháp xây\_dựng lòng tin sẵn có và được thiết\_kế để đảm\_bảo việc tuẫn\_trá dọc đường kiểm\_soát thực\_tế ( hay còn gọi là đường biên\_giới chưa phân định ) không leo\_thang thành các cuộc giao\_tranh . Thủ\_tướng Ấn Độ\_Manmohan\_Singh cũng cho rằng , " thỏa\_thuận này sẽ bổ Sung thêm các công\_cụ hiện có để bảo\_dảm hòa\_bình , ổn định ở khu\_vực biên\_giới ". Theo thỏa\_thuận mới được ký\_kết , 2 nước sẽ thông\_báo cho nhau về các cuộc tuẫn\_trá dọc biên\_giới , đảm\_bảo các cuộc tuẫn\_trá này không " bám đuôi " nhau để giảm\_thiểu khă\_năng đối\_dấu và sẽ thực\_hiện " kiêm\_chế tối\_da " nếu 2 bên đối\_mặt nhau ở những khu\_vực có đường\_bien chưa rõ ràng . Quân\_đội 2 nước được bố\_trí suốt dọc 4.000 km chiều dài biên\_giới từ cao\_nguyên Ladakh ở phía Tây đến các khu rừng thuộc Arunachal\_Pradesh ở phía Đông . Các quan\_chức của Trung\_Quốc và Ấn Độ cũng đã đồng\_y sẽ thiết lập đường\_dây\_nóng để có\_thể bám sát giải\_quyết các vấn đề một\_cách nhanh\_chóng và kịp\_thời . Thỏa\_thuận về vấn đề biên\_giới mà các nhà ngoại\_giao 2 nước đã được gấp\_rút hoàn\_thành trước khi Thủ\_tướng Ấn Độ\_Manmohan\_Singh sang thăm Trung\_Quốc . Theo đánh\_giá của các chuyên\_gia phân\_tích , đây chỉ là một bước\_tiến nhỏ trong mối quan\_hệ phức\_tap Trung - Ấn . Tháng 5/2013 , quân\_đội 2 nước đã có 3 tuần căng\_thẳng ở phía Tây dãy Himalaya sau khi quân\_đội Trung\_Quốc dựng trại sâu bên trong lãnh\_thổ Ấn Độ khoảng 10km . Sự việc này gây ra lùn\_sóng phản ứng mạnh\_mẽ của người\_dân Ấn Độ , thậm chí họ còn kêu\_gọi Chính\_phủ phải có hành\_dòng cứng\_rắn đối\_với Trung\_Quốc . Tuy\_nhiên , Trung\_Quốc phủ\_nhận việc quân\_đội nước này đã xâm\_nhập lãnh\_thổ Ấn Độ . Trước chuyến thăm Trung\_Quốc của Thủ\_tướng Singh , đầu tháng này , quan\_hệ 2 nước một lần nữa lại nổi sóng khi Trung\_Quốc chỉ đồng\_y cấp thi\_thực rời cho 2 cung thủ Ấn Độ đến từ vùng tranh\_chấp Arunachal\_Pradesh khi 2 vận động\_viên tham\_gia một giải bắn cung ở Trung\_Quốc . Để đáp trả , New\_Delhi đã quyết định tạm ngưng thỏa\_thuận nói\_lòng thi\_thực nhập\_cánh của Ấn Độ với Trung\_Quốc vào phút chót . Tuy\_nhiên , ông Lý\_Khắc\_Cường đã tìm cách xoa\_diu căng\_thẳng khi tuyên\_bố : " Trung\_Quốc và Ấn Độ là 2 nền văn\_minh lâu\_dời ... Chính\_phủ của 2 nước chúng\_tôi có khă\_năng quản\_lý những bất đồng ở khu\_vực biên\_giới , để vấn đề này không làm ảnh\_hưởng đến lợi\_ich tổng\_thể trong mối quan\_hệ song\_phương giữa 2 nước chúng\_tôi " . Theo thỏa\_thuận mới được ký\_kết , 2 nước sẽ thông\_báo cho nhau về các cuộc tuẫn\_trá dọc biên\_giới , đảm\_bảo các cuộc tuẫn\_trá này không " bám đuôi " nhau để giảm\_thiểu khă\_n năng đối\_dấu và sẽ thực\_hiện " kiêm\_chế tối\_da " nếu 2 bên đối\_mặt nhau ở những khu\_vực có đường\_bien chưa rõ ràng . Trong lịch\_sử , 2 nước đã có một cuộc chiến\_tranh biên\_giới ngắn ngày vào năm 1962 khiến mối quan\_hệ sau đó giữa Trung\_Quốc và Ấn Độ luôn ở trong trạng\_thái ngõ\_vực . Tháng 5/2013 , quân\_đội 2 nước đã có 3 tuần căng\_thẳng ở phía Tây dãy Himalaya sau khi quân\_đội Trung\_Quốc dựng trại sâu bên trong lãnh\_thổ Ấn Độ khoảng 10 km . Trong khi đó , Ấn Độ cho rằng , Trung\_Quốc chiếm 38.000 km<sup>2</sup> lãnh\_thổ của nước này trên cao\_nguyên Aksai\_Chin ở phía tây Himalaya .

## VĂN BẢN TÓM TẮT

Tuy\_nhiên , ông Lý\_Khắc\_Cường đã tìm cách xoa\_diu căng\_thẳng khi tuyên\_bố : " Trung\_Quốc và Ấn Độ là 2 nền văn\_minh lâu\_dời ... Chính\_phủ của 2 nước chúng\_tôi có khă\_n năng quản\_lý những bất đồng ở khu\_vực biên\_giới , để vấn đề này không làm ảnh\_hưởng đến lợi\_ich tổng\_thể trong mối quan\_hệ song\_phương giữa 2 nước chúng\_tôi " . Theo thỏa\_thuận mới được ký\_kết , 2 nước sẽ thông\_báo cho nhau về các cuộc tuẫn\_trá dọc biên\_giới , đảm\_bảo các cuộc tuẫn\_trá này không " bám đuôi " nhau để giảm\_thiểu khă\_n năng đối\_dấu và sẽ thực\_hiện " kiêm\_chế tối\_da " nếu 2 bên đối\_mặt nhau ở những khu\_vực có đường\_bien chưa rõ ràng . Trong lịch\_sử , 2 nước đã có một cuộc chiến\_tranh biên\_giới ngắn ngày vào năm 1962 khiến mối quan\_hệ sau đó giữa Trung\_Quốc và Ấn Độ luôn ở trong trạng\_thái ngõ\_vực . Tháng 5/2013 , quân\_đội 2 nước đã có 3 tuần căng\_thẳng ở phía Tây dãy Himalaya sau khi quân\_đội Trung\_Quốc dựng trại sâu bên trong lãnh\_thổ Ấn Độ khoảng 10 km . Trong khi đó , Ấn Độ cho rằng , Trung\_Quốc chiếm 38.000 km<sup>2</sup> lãnh\_thổ của nước này trên cao\_nguyên Aksai\_Chin ở phía tây Himalaya .

Hình 3.17 Test th2

## Kết quả TH3 Lấy văn bản từ video để test

**VĂN BẢN GỐC**

tác hại kinh hoàng của thuốc lá điện tử có thể nhiều người chưa biết hút thuốc lá điện tử không độc hại hút thuốc lá điện tử cho Sành điệu là tâm lý của nhiều bạn trẻ mới lớn nhưng thực tế không ít trường hợp đã bị ngộ độc ma túy tổn thương nội tạng và sức khỏe vong những điều thuốc lá điện tử nhiều màu sắc hình thức nhỏ gọn bắt mắt hương thơm hấp dẫn như keo trái cây được giới trẻ ưa chuộng nhất là học sinh sinh viên đi kèm Đó là những lời quảng cáo không gây hại văn hóa thuốc lá lành mạnh sành điệu thuốc lá thế hệ mới đã đánh trúng tâm lý thích thể hiện cái tôi độ chịu chơi thích khám phá của tuổi mới lớn tiền sỹ Nguyễn Trung Nguyên giám đốc trung tâm chống độc bệnh viện Bạch Mai cho biết vài năm trở lại đây hầu như tuần nào trung tâm cũng tiếp nhận một vài ca ngộ độc sau khi hút thuốc lá điện tử trường hợp bệnh nhân nhập viện trong tình trạng ngộ độc rất nặng lần quay ra tiền tới sự bất ngờ co giật suốt từ vòng cổ trường hợp bệnh nhân nhẹ thì đến viện trong tình trạng ngơ ngác rồi loạn tâm thần hoang tưởng ào giác kích thích lờ đờ tần thương các cơ quan nội tạng đổi tượng sử dụng thuốc lá điện tử hầu hết là người trẻ và có nhiều trường hợp là học sinh Trung học phổ thông mới Đây khoa chống độc đã tiếp nhận một trường hợp thanh niên 23 tuổi bị ngộ độc thuốc lá điện tử bệnh nhân nhập viện trong tình trạng sùi bọt mép co giật hồn mê loạn thần sau khi xét nghiệm các bác sĩ tìm thấy ba loại chất ma túy có trong thuốc lá điện tử có rất nhiều loại chất khi bệnh nhân hit phải sẽ kích thích kinh tim mạch rất mạnh đây là những loại ma túy cực mạnh hoàn toàn mới điều này cho thấy các chất ma túy thế hệ mới thay đổi hàng ngày và dưới hình thức sử dụng được thay đổi nhiều dạng khác nhau có thể hút trực tiếp có thể được trộn lẫn vào hóa chất có trong thuốc lá điện tử trong thuốc lá điện tử có chất nicotine liều lượng có thể rất thấp hoặc không có Tuy nhiên nó lại có chất độc gây co mạch tổn thương cho cơ thể bạn đã nói bản chất của nicotine là gây nghiện người chưa sử dụng thuốc lá bao giờ nếu sử dụng thuốc lá điện tử sẽ gây nghiện nicotine từ đó di chuyển sang nghiện thuốc lá thuốc lá thông thường hiện có khoảng 15,500 loại hương liệu được sử dụng trong các sản phẩm thuốc lá điện tử trong đó rất nhiều các loại hương liệu được xem là chất độc và chưa được đánh giá toàn diện về mức độ gây hại đối với sức khỏe tình đầu chứa trong thuốc lá điện tử là một hỗn hợp bao gồm nicotine có chất tạo mùi dung môi và giúp phu gia khác khi hút thuốc lá điện tử chất lỏng này sẽ được đốt nóng và tạo khói với nhiều mùi thơm khác nhau bác sĩ Nguyễn cho biết đã có những trường hợp gây tổn thương phổi cấp do Vitamin E có trong thuốc lá điện tử Lý do là vitamin E gây đốt cháy nén chất độc gây tổn thương phổi hàng nghìn người đã tử vong vì loại chất này ở Việt Nam cũng đã phát hiện có vitamin E trong thuốc lá điện tử chất này kết hợp với nhiều hóa chất khác nữa gây nên một loạt bệnh mới nổi mà không biết điều này tạo nên gánh nặng với xã hội và cả hệ thống y tế thuốc lá điện tử xuất hiện rần rần trên thị trường hiện nay giá dao động từ vài chục nghìn đến triệu đồng nếu bắt cứ ai bắt cứ lứa tuổi nào cũng có thể mua được thực tế hiện nay thuốc lá điện tử vẫn chưa được cấp phép kinh doanh Tuy nhiên trên trang mạng loại thuốc này vẫn được bán rầm rộ lan truyền trên thị trường nó có thể gây tác động sớm đối với sức khỏe Hoặc gây bệnh phổi kẽ Nếu so với căn bệnh ung thư phổi bệnh phổi kẽ tiến triển nhanh và tỉ lệ sống Ngoài ra thuốc lá điện tử cũng không có công dụng cai nghiện Thuốc Lá Điều thông thường Hằng ngày ký Kênh alo bác sĩ video để nhận thêm thông tin mới nhất về y tế sức khỏe

**VĂN BẢN TÓM TẮT**

Sau đó, bệnh viện đang ở trong tình trạng hoang mang của các rỗi loạn tâm thần, về rối loạn tâm thần hoang tưởng ào tưống kích thích các điện tử gây ra chủ yếu là những người trẻ tuổi và có nhiều trường hợp học sinh trung học gần đây đã nhận được một trường hợp 23 tuổi - Những người trẻ tuổi bị ngộ độc thuốc lá điện tử, bệnh nhân nhập viện trong tình trạng tạo bọt mép của cơ co giật của hồn mê tâm thần sau khi các bác sĩ tìm thấy ba loại thuốc trong thuốc lá điện tử, càng bị nhiễm độc sau khi hút thuốc. Việc sử dụng các thay đổi khác nhau có thể được hút trực tiếp có thể được trộn vào hóa chất trong thuốc lá điện tử trong thuốc lá điện tử với liều nicotine có thể rất thấp hoặc không, nhưng đó là một chất độc hại Điều đó khiến các tổn thương mạch máu cho cơ thể bạn nói rằng bản chất của Nicotine gây nghiện cho những người chưa bao giờ sử dụng thuốc lá nếu sử dụng E-cigarettes sẽ gây nghiện cho nicotine, do đó chuyển sang hút thuốc. Hàng ngàn người đã chết vì chất này ở Việt Nam cũng đã phát hiện ra vitamin E trong thuốc lá điện tử kết hợp với nhiều hóa chất khác gây ra một loạt các bệnh mới nổi mà không biết điều này tạo ra gánh nặng cho xã hội và hệ thống y tế thuốc lá. Lào thường có khoảng 15.500 loại hương liệu được sử dụng trong các sản phẩm thuốc lá điện tử, trong đó nhiều loại mùi được coi là độc tố và chưa được đánh giá toàn diện về mặt tác hại đối với sức khỏe. Bệnh viện cho biết trong vài năm qua, gần như mỗi tuần, trung tâm đã nhận được một vài nhà máy ngộ độc sau khi hút thuốc lá điện tử.

Hình 3.17 Test th3

### 3.4.4. Nhận xét về kết quả thu được:

Nhận xét về kết quả thu được của từng trường hợp

## 1. Trường hợp tóm tắt văn bản có sẵn copy vào



Hình 3.19 Kết quả th1

Trong quá trình thực hiện tóm tắt văn bản bằng Python, kết quả thu được đã phản ánh một cách rõ ràng và tương đối đầy đủ các ý chính của văn bản gốc. Cụ thể, các nội dung về địa lý, văn hóa và các đặc trưng du lịch của tỉnh Nam Định trong văn bản gốc đã được giữ lại trong bản tóm tắt, đảm bảo tính chính xác và mạch lạc.

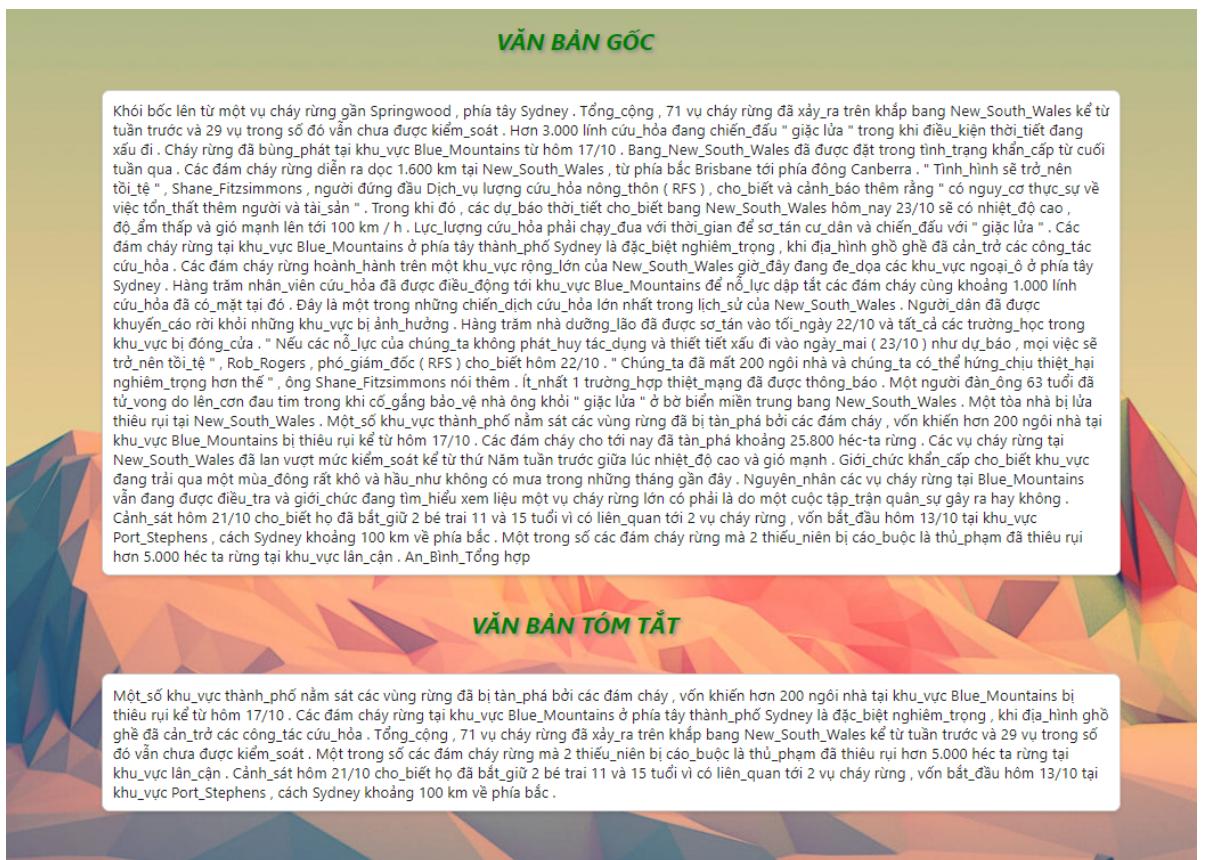
Tuy nhiên, kết quả tóm tắt vẫn còn một số điểm cần cải thiện như sau:

**Tính cô đọng:** Mặc dù văn bản đã được rút gọn, nhưng vẫn còn một số câu văn không cần thiết có thể lược bỏ để giúp đoạn tóm tắt ngắn gọn và súc tích hơn.

- Tính toàn diện:** Một vài thông tin chi tiết quan trọng về lễ hội và ẩm thực đặc trưng của địa phương chưa được đề cập đầy đủ trong bản tóm tắt. Điều này có thể làm giảm đi tính toàn diện của nội dung.

- Tính liên quan: Một số câu mở đầu trong phần tóm tắt chưa thực sự liên quan chặt chẽ với nội dung chính của văn bản gốc, khiến đoạn tóm tắt có phần lan man.
- Về tổng thể, thuật toán tóm tắt văn bản đã thực hiện tốt nhiệm vụ cung cấp thông tin, tuy nhiên cần cải thiện để đảm bảo kết quả tóm tắt ngắn gọn, chính xác và bao quát hơn, đáp ứng tốt hơn các yêu cầu đặt ra trong đồ án.

## 2. Trường hợp tóm tắt văn bản có sẵn từ text



Hình 3.20 Kết quả th2

Trong quá trình thực hiện tóm tắt văn bản bằng Python, kết quả thu được đã phản ánh một cách rõ ràng và tương đối đầy đủ các ý chính của văn bản gốc. Cụ thể về các vụ cháy rừng tại khu vực New South Wales, thiệt hại

về người và tài sản, cũng như các nỗ lực của lực lượng cứu hỏa tại khu vực Blue Mountains. Tuy nhiên, vẫn có một số điểm cần xem xét như sau:

- Độ chính xác nội dung: Kết quả tóm tắt đã giữ lại các ý chính về vụ cháy rừng, nhưng một số chi tiết quan trọng, chẳng hạn như thông tin về tình trạng thời tiết khắc nghiệt và những khó khăn trong công tác chữa cháy, không được thể hiện đầy đủ trong bản tóm tắt.
- Tính cô đọng: Văn bản đã được rút gọn đáng kể, tuy nhiên, có thể làm cô đọng thêm ở một số câu, đặc biệt là những phần nhắc đến chi tiết về việc bắt giữ các cá nhân liên quan đến vụ cháy rừng, để tránh làm tóm tắt trở nên rườm rà.
- Tính mạch lạc và dễ hiểu: Phần tóm tắt khá mạch lạc và dễ hiểu. Tuy nhiên, có một vài câu có thể được viết ngắn gọn hơn để đảm bảo người đọc nắm bắt thông tin nhanh chóng và chính xác.
- Độ bao quát thông tin: Mặc dù bản tóm tắt đã bao quát một số thông tin chính, nhưng các nội dung về mức độ nghiêm trọng của các đám cháy rừng, quy mô thiệt hại, và sự hợp tác giữa các cơ quan chức năng trong công tác cứu hỏa chưa được đề cập đầy đủ. Điều này có thể làm giảm tính toàn diện của bản tóm tắt.

Kết luận, kết quả tóm tắt đã đáp ứng một phần yêu cầu nhưng cần cải thiện để bao quát đầy đủ nội dung và đảm bảo tính cô đọng, dễ hiểu hơn. Điều này sẽ giúp người đọc dễ dàng nắm bắt các điểm chính mà không bỏ lỡ những thông tin quan trọng.

### 3. Trường hợp tóm tắt văn bản từ video đăng tải lên

**VĂN BẢN GỐC**

kể từ năm 1980 đến nay thi thể thao Việt Nam đã tham dự Tổng cộng là 10 km/h mới dành được tổng cộng có 5 huy chương thời Trong đó có một huy chương vàng Ba huy chương bạc và 1 huy chương đồng nên việc chúng ta mòn mỏi chờ đợi tấm huy chương ở Olympic này cũng là một điều khá là bình thường bởi kể từ khi xạ thủ Hoàng Xuân Vinh mang về tấm huy chương vàng Olympic đầu tiên và duy nhất cho tới nay thi kỳ Olympic này những niềm hi vọng cũng đều mang tên Vinh cả đầu tiên chính là xạ thủ Trịnh Thu Vinh hẵn quý vị vẫn còn nhớ ở chung kết nội dung 10m súng ngắn thử Bình đã vươn mặt tấm huy chương đồng cách có vài điểm thôi Và giờ thì có đang tràn đầy cơ hội có thể mang về tấm huy chương đầu tiên cho thể thao Việt Nam khi đã có mặt ở chung kết ở nội dung 25m súng ngắn thể thao ở vòng loại nội dung 25m súng ngắn thể thao nữ xạ thủ Trịnh Thu Vinh xếp thứ 4 trên tổng số 40 vận động viên tham dự và giành quyền vào chung kết trong số 8 xạ thủ sẽ Tranh Tài ở chung kết đáng chú ý có Yang Jin hạng 2 thế giới Kalin hạng ba thế giới Cùng với đó là veronica và janus những người đang đứng trong top 10 thế giới phán tích chung kết ngoài phong độ thi tâm lý thi đấu cũng là yếu tố quyết định tới việc thu Vinh có thể giành huy chương hay không phong độ của các vận động viên bắn súng thường khá ổn định Đôi khi họ thể hiện phong độ rất tốt nhưng chỉ một sai lầm nhỏ thôi cũng có thể ảnh hưởng đến kết quả chung cuộc thi Vinh Cần cải thiện là về tâm lý thi đấu như vậy nếu có một Khối Đầu Tốt cùng một tâm lý ổn định thu Vinh hoàn toàn có thể giành huy chương ở thế vận hội lần này phải thi chung kết nội dung 205m súng ngắn thể thao nữ sẽ diễn ra vào lúc 14:30 hôm nay niềm hy vọng tiếp là một cái tên Vinh nữa đó chính là Trịnh Văn Vinh ở môn cù tạ thời điểm này thi Vinh đã có mặt tại Paris để chuẩn bị Tranh Tài ở hạng 61kg trước đó thi Trịnh Văn Vinh đã có chuyến tập huấn hơn 40 ngày tại Trung Quốc cộng thêm với 5 giải Vòng loại đủ giúp cho lực sĩ này tự tin trước kỳ Olympic đầu tiên trong sự nghiệp Nếu như ở các giải khát Vinh chỉ được sang trước 12 ngày thi tại Olympic Paris Vinh và hai huấn luyện viên được thăng chức cả tuần ở Paris để chuẩn bị vận động viên được sang đây là 7 ngày Đây là một cải thiện lớn cho người ta để làm quen được cái mùi giòi và cái thời tiết và ăn uống sinh hoạt mong rằng đây là kỳ Olympic để vận động viên phát huy được cái tài năng vốn có của mình huy chương như cù tạ việc ăn uống đóng vai trò vô cùng quan trọng vì thế trước việc đồ ăn ở Olympic chuyên về các món chay và rau bồ Vinh và 301 phải có những đổi sách vừa nghe ngóng được cái tin như vậy thi bắn luyện và vận động viên sẽ chuẩn bị các thực phẩm chức năng và chưa đồ ăn cho vận động viên từ nhà mang sang để cho đảm bảo an toàn thi được vận động viên Thể thi khả năng này em sẽ phải chuẩn bị thêm một số thịt thịt hộp hay là cái đồ ăn nhanh và thời cũng sẽ chuẩn bị những cái thức ăn như là đồ ăn nhanh cháo các thứ thi không bị đồ ăn hồn rượu Vinh sẽ có hại thay theo để hỗ trợ về tâm lý phục hồi và chỉ đạo chiến thuật trên sân tất cả với mục tiêu lớn nhất là giành huy chương của hảng 61kg và cử tạ từng mang về hai huy chương cho thể thao Việt Nam ở đấu trường Olympic đó là huy chương bạc của vận động viên Hoàng Anh Tuấn và huy chương đồng của Trần Lê Quốc Toản năm nay thi hi vọng huy chương lại đồn áp lực lên đôi vai của Trịnh Văn Vinh với mức tổng cử tốt nhất là 294 kg đã giúp mình có mặt tại Olympic Paris nhưng anh chỉ đứng thứ 9 trên bảng xếp hạng ở hạng 61kg cùng cạnh tranh ở hạng cân này còn có các lực sĩ rất mạnh đến từ Trung Quốc Indonesia Thái Lan Mỹ cơ hội huy chương của mình là khá sáng nếu đạt được tổng cử 294 kg anh có thể tiếp cận với tấm huy chương đồng mà ig solga được tại Olympic Tokyo 2020 Violympic này Vinh sẽ gặp lại người quen cũ đó là Eco roe từng hai lần giành huy chương vàng Olympic và anh cũng là đại tướng của Vinh tại SEA Games 2017 năm đó thi Vinh đã phá kỷ lục với tổng cử là 307 kg và nếu lập lại được kỷ tích này ở Olympic năm nay cơ hội giành huy chương của mình là rất sáng bởi huy chương bạc ở hạng cân này tại Olympic Tokyo tổng cử Có 302 kg mà thôi

**VĂN BẢN TÓM TẮT**

Giành được huy chương của anh ấy rất sáng vì huy chương bạc trong hạng cân này tại Olympictokyo Total chỉ có 302 kg Với 5 vòng đủ điều kiện , để giúp vận động viên này tự tin trước Thế vận hội Olympic đầu tiên trong sự nghiệp, nếu trong đồ uống Vinh , Vinh và hai huấn luyện viên được thăng cấp cả hai chương trình khuyến mãi trong tuần ở Paris để chuẩn bị cho các vận động viên đến đây là 7 ngày . Vì vậy , họ làm việc , chúng tôi đang chờ đợi huy chương trong Olympic này là khá bình thường , bởi vì xạ thủ Hoang Xuan Vinh đã mang lại huy chương vàng Olympic đầu tiên và duy nhất cho đến nay tên của Vinh , xạ thủ Trịnh Thu Vinh - bạn vẫn phải nhớ trận chung kết của nội dung 10m của khẩu súng lực của khẩu súng lực của những người trúng phạt bị mất huy chương đồng . Với một vài điểm , và bây giờ có ấy đang tràn đầy cơ hội để có thể mang lại huy chương đầu tiên cho các môn thể thao Việt Nam khi đó là trận chung kết trong nội dung 25m của súng thể thao ở vòng vòng 25m của nữ xạ thủ thể thao Trịnh . Thu Vinh ranked thứ 4 trong số 40 vận động viên tham dự và giành quyền trận chung kết trong số 8 xạ thủ sẽ thi đấu trong trận chung kết đáng chú ý với Yang Jin , hạng 2 thế giới Kalin trên thế giới , cùng với Veronica và Veronica và Top 10 của thế giới phán tích trận chung kết , ngoài hình thức của họ , tâm lý cạnh tranh cũng là yếu tố quyết định của Thu Vinh , có nên giành huy chương hay không . Hãy liên quan đến tâm lý phục hồi và chiến thuật trên sân , tất cả đều có mục tiêu lớn nhất là giành huy chương 61kg và cử tạ một lần mang lại hai huy chương cho các môn thể thao Việt Nam trong OlympicĐấu trường .

Hình 3.21 Kết quả th3

Trong quá trình thực hiện tóm tắt văn bản bằng Python, kết quả thu được đã phản ánh một cách rõ ràng và tương đối đầy đủ các ý chính của văn bản gốc. Phân tóm tắt văn bản đã phản ánh các thông tin chính từ văn bản gốc, bao gồm các thành tích của Trịnh Văn Vinh trong thể thao, quá trình thi đấu Olympic, và các chi tiết về hành trình giành huy chương của anh. Tuy nhiên, vẫn có một số điểm cần xem xét như sau:

- Độ chính xác nội dung: Kết quả tóm tắt đã bao quát được những ý chính của văn bản gốc, bao gồm thông tin về sự nghiệp thể thao của Trịnh Văn Vinh và các cuộc thi quốc tế mà anh tham gia. Tuy nhiên, một số chi tiết như sự kiện cụ thể và các thành tích khác ngoài việc giành huy chương có thể bị lược bỏ.
- Tính cô đọng: Phần tóm tắt đã được rút gọn nhưng vẫn còn một số câu dài và không cần thiết. Có thể làm cô đọng hơn nữa để văn bản trở nên ngắn gọn và dễ theo dõi hơn.
- Tính mạch lạc và dễ hiểu: Kết quả tóm tắt có tính mạch lạc và tương đối dễ hiểu. Tuy nhiên, cần chỉnh sửa lại một số câu để người đọc dễ dàng nắm bắt nội dung chính mà không phải đọc qua nhiều chi tiết phụ.
- Độ bao quát thông tin: Phần tóm tắt đã giữ được các nội dung quan trọng, nhưng chưa đề cập đầy đủ đến những thành tích khác của Trịnh Văn Vinh, chẳng hạn như những giải đấu quốc tế khác mà anh đã tham gia và đóng góp cho thể thao Việt Nam. Điều này có thể làm giảm tính toàn diện của bản tóm tắt.

Kết luận, kết quả tóm tắt đã đạt được một phần yêu cầu, tuy nhiên cần cải thiện thêm để đảm bảo tính cô đọng, dễ hiểu và bao quát đủ nội dung hơn. Điều này sẽ giúp người đọc nắm bắt các thông tin quan trọng mà không bỏ lỡ những chi tiết cần thiết.

### 3.5. Hạn chế và hướng phát triển

#### 3.5.1. Hạn chế của nghiên cứu:

Phụ thuộc vào chất lượng dữ liệu đầu vào:

- Nghiên cứu có thể bị ảnh hưởng bởi chất lượng của dữ liệu đầu vào. Nếu văn bản đầu vào chứa nhiều nhiễu hoặc không được chuẩn hóa đúng cách, kết quả tóm tắt có thể không chính xác hoặc thiếu sót những thông tin quan trọng.
- Giải Pháp: Tăng cường quá trình tiền xử lý dữ liệu bằng cách sử dụng các phương pháp nâng cao như loại bỏ nhiễu, chuẩn hóa văn bản để cải thiện chất lượng đầu vào trước khi áp dụng thuật toán tóm tắt.

Giới hạn của textrank và phương pháp tính độ tương đồng:

- TextRank dựa trên việc đếm số lượng từ chung để tính toán mức độ tương đồng giữa các câu, vì vậy nó có những giới hạn nhất định trong việc hiểu ngữ cảnh và ý nghĩa sâu sắc của văn bản. Điều này có thể dẫn đến việc bỏ sót các yếu tố ngữ nghĩa phức tạp trong văn bản.
- Giải Pháp: Khám phá và tích hợp các mô hình ngôn ngữ sâu hơn như BERT hoặc Transformer, giúp nắm bắt ngữ cảnh và ý nghĩa của văn bản một cách hiệu quả hơn.

Scalability và hiệu suất:

- Khi áp dụng thuật toán TextRank trên các bộ dữ liệu văn bản lớn hoặc phức tạp, hiệu suất và thời gian xử lý có thể trở thành một thách thức, đặc biệt là với các tài nguyên tính toán hạn chế.
- Giải Pháp: Tối ưu hóa thuật toán, sử dụng các phương pháp tính toán phân tán hoặc các hệ thống lưu trữ và xử lý dữ liệu lớn để cải thiện khả năng mở rộng và hiệu suất khi làm việc với các tập dữ liệu lớn.

### 3.5.2. Hướng phát triển tiềm năng:

Tích hợp các mô hình ngôn ngữ tiên tiến:

- Nghiên cứu có thể được mở rộng bằng cách tích hợp các mô hình ngôn ngữ tiên tiến như BERT hoặc GPT, để cải thiện khả năng hiểu và xử lý ngữ cảnh của văn bản, từ đó nâng cao chất lượng tóm tắt.
- **Ưu Điểm:** Những mô hình ngôn ngữ này sẽ giúp nắm bắt các nghĩa phức tạp và ngữ cảnh tốt hơn, từ đó tạo ra các bản tóm tắt có độ chính xác và sự mạch lạc cao hơn.

Khám phá phương pháp tóm tắt đa ngôn ngữ:

- Mở rộng nghiên cứu để hỗ trợ và đánh giá hiệu quả của thuật toán TextRank trên nhiều ngôn ngữ khác nhau, đặc biệt là những ngôn ngữ có cấu trúc ngữ pháp phức tạp.
- **Ưu Điểm:** Tăng cường khả năng áp dụng của thuật toán trên các tập dữ liệu đa ngôn ngữ, phục vụ cho các ứng dụng toàn cầu và ngữ cảnh văn hóa đa dạng.

Tối ưu hóa và mở rộng quy mô:

- Tập trung vào việc tối ưu hóa thuật toán TextRank và mở rộng quy mô, đảm bảo hiệu suất khi áp dụng trên các bộ dữ liệu lớn. Điều này có thể đạt được bằng cách tối ưu hóa quy trình tính toán và sử dụng các công nghệ hiện đại.
- **Ưu Điểm:** Cải thiện khả năng mở rộng và hiệu suất của thuật toán, giúp nó phù hợp hơn với các ứng dụng thực tế và dữ liệu lớn.

## KẾT LUẬN

Trong quá trình thực hiện đề tài "Tóm tắt văn bản tự động và ứng dụng", em đã tiến hành nghiên cứu và triển khai các phương pháp tóm tắt văn bản tự động dựa trên thuật toán TextRank và PageRank. Dưới đây là một số tổng kết mà em rút ra:

Em đã phát triển mô hình tóm tắt văn bản tự động, sử dụng thuật toán TextRank, cho thấy khả năng hiệu quả trong việc xử lý và tóm tắt các văn bản lớn, mang lại kết quả chính xác và hữu ích.

Mô hình được huấn luyện với dữ liệu đa dạng, giúp nó tổng quát hóa và áp dụng được trên nhiều loại văn bản, từ đó đảm bảo tính linh hoạt và độ chính xác trong thực tế.

Mặc dù mô hình đã cho kết quả tích cực, nhưng vẫn tồn tại một số hạn chế cần khắc phục, đặc biệt là trong các tình huống văn bản có sự phức tạp cao hoặc cấu trúc ngôn ngữ đặc biệt.

Đề tài cũng đưa ra thách thức về việc cần thiết lập thêm các thử nghiệm trên tập dữ liệu lớn và phức tạp hơn, cũng như đánh giá hiệu suất mô hình trong các ứng dụng thực tế.

Em hy vọng rằng kết quả nghiên cứu này sẽ góp phần cải thiện công cụ xử lý ngôn ngữ tự nhiên và mở ra hướng phát triển mới trong lĩnh vực trí tuệ nhân tạo và xử lý dữ liệu văn bản.

# TÀI LIỆU THAM KHẢO

## **1. Tác phẩm sách:**

- [1]"Xử lý ngôn ngữ và giọng nói" của Daniel Jurafsky và James H. Martin
- [2]"Xử lý ngôn ngữ tự nhiên bằng Python" của Steven Bird, Ewan Klein và Edward Loper
- [3]"Nền tảng của Xử lý ngôn ngữ tự nhiên theo thống kê" của Christopher D. Manning và Hinrich Schütze
- [4]"Học sâu để xử lý ngôn ngữ tự nhiên" của Zhiheng Huang, Wei Xu và Kai Yu
- [5]"Xử lý ngôn ngữ tự nhiên trong thực tế" của Hobson Lane, Hannes Hapke và Cole Howard

## **2. Website:**

- [1] Flask Documentation. Flask Documentation, Flask API Documentation, available at:<https://flask.palletsprojects.com/en/2.0.x/>.
- [2] Werkzeug Documentation. Werkzeug – The Python WSGI Utility Library, available at: <https://werkzeug.palletsprojects.com/en/2.0.x/utils/>.
- [3] MoviePy Documentation. MoviePy Video Editing with Python, available at: <https://zulko.github.io/moviepy/>.
- [4] Requests Documentation. Requests: HTTP for Humans, available at: <https://docs.python-requests.org/en/latest/>.
- [5] Underthesea Documentation. Xử lý ngôn ngữ tự nhiên tiếng Việt với Underthesea, available at:<https://github.com/undertheseanlp/underthesea>.

[6] SpeechRecognition Documentation. SpeechRecognition Python Library Documentation, available at:  
<https://pypi.org/project/SpeechRecognition/>.

[7] Googletrans Documentation. Googletrans Python API Documentation, available at:  
<https://py-googletrans.readthedocs.io/en/latest/>.

[8] NetworkX Documentation. NetworkX Library for Graphs and Networks, available at:  
<https://networkx.github.io/documentation/stable/>.

### **3. Tài liệu học thuật trực tuyến (ebook, học liệu trực tuyến):**

[1]: Nguyễn Viết Hạnh (2018), Nghiên cứu tóm tắt văn bản tự động và ứng dụng, Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.