

**-----------------------------------**

**NGUYỄN ĐÌNH SƠN**

PHÁT TRIỂN MÔ HÌNH TEXT MINING DỰA TRÊN

KỸ THUẬT MACHINE LEARNING CHO TÓM TẮT

VĂN BẢN TIẾNG VIỆT

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

**TP. HỒ CHÍ MINH – NĂM 2024**

**--------------------------------------**



**NGUYỄN ĐÌNH SƠN**

PHÁT TRIỂN MÔ HÌNH TEXT MINING DỰA TRÊN

KỸ THUẬT MACHINE LEARNING CHO TÓM TẮT

VĂN BẢN TIẾNG VIỆT

|  |  |
| --- | --- |
| **Chuyên ngành:** | **Hệ thống thông tin** |
| **Mã số:** | **8.48.01.04** |

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

|  |
| --- |
| NGƯỜI HƯỚNG DẪN KHOA HỌC: |
| **TS. TÂN HẠNH** |

**TP. HỒ CHÍ MINH – NĂM 2024**

# LỜI CAM ĐOAN

Tôi cam đoan rằng đề án tốt nghiệp thạc sĩ: ***“Phát triển mô hình Text Mining dựa trên kĩ thuật Machine Learning cho tóm tắt văn bản tiếng Việt”*** là công trình nghiên cứu của chính tôi.

Các số liệu, kết quả nêu trong đề án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Không có sản phẩm/nghiên cứu nào của người khác được sử dụng trong đề án này mà không được trích dẫn theo đúng quy định.

TP. Hồ Chí Minh, ngày … tháng … năm 2024

|  |  |  |  |
| --- | --- | --- | --- |
| **Học viên thực hiện đề án** | | | |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **Nguyễn Đình Sơn** | | | |

# LỜI CẢM ƠN

Trong suốt quá trình học tập và nghiên cứu thực hiện đề án tốt nghiệp thạc sĩ, ngoài nỗ lực của bản thân, tôi đã nhận được sự hướng dẫn nhiệt tình quý báu của quý Thầy Cô, cùng với sự động viên và ủng hộ của gia đình, bạn bè và đồng nghiệp. Với lòng kính trọng và biết ơn sâu sắc, tôi xin gửi lời cảm ơn chân thành tới:

**Thầy TS. Tân Hạnh**, người Thầy kính yêu đã hết lòng giúp đỡ, hướng dẫn, động viên, tạo điều kiện cho tôi trong suốt quá trình thực hiện và hoàn thành đề án tốt nghiệp thạc sĩ.

Ban Giám Đốc, Phòng đào tạo sau đại học và quý Thầy Cô đã tạo mọi điều kiện thuận lợi giúp tôi hoàn thành đề án.

Tôi xin chân thành cảm ơn gia đình, bạn bè, đồng nghiệp trong cơ quan đã động viên, hỗ trợ tôi trong lúc khó khăn để tôi có thể học tập và hoàn thành đề án. Mặc dù đã có nhiều cố gắng, nỗ lực, nhưng do thời gian và kinh nghiệm nghiên cứu khoa học còn hạn chế nên không thể tránh khỏi những thiếu sót. Tôi rất mong nhận được sự góp ý của quý Thầy Cô cùng bạn bè đồng nghiệp để kiến thức của tôi ngày một hoàn thiện hơn.

Xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày … tháng … năm 2024

|  |  |  |  |
| --- | --- | --- | --- |
| **Học viên thực hiện đề án** | | | |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **Nguyễn Đình Sơn** | | | |

# DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **VIẾT TẮT** | **TIẾNG ANH** | **TIẾNG VIỆT** |
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| TF-IDF | Term Frequency-Inverse Document Frequency | Tần suất xuất hiện - Tần suất nghịch đảo tài liệu |
| TM | Text Minig | Khai phá văn bản |
| TDM | Text Data Mining | Khai phá dữ liệu văn bản |
| KDT | Discovery in Textual Databases | Khai phá tri thức trong cơ sở dữ liệu văn bản |
| LSA | Latent Semantic Analysis | Phân tích ngữ nghĩa ẩn |
| LSI | Latent Semantic Indexing | Lập chỉ mục ngữ nghĩa ẩn |
| SVD | Singular Value Decomposition | Phân tích giá trị kỳ dị |
| IR | Information Retrieval | Truy hồi thông tin |
| CLIR | Cross-Language Information Retrieval | Truy hồi thông tin đa ngôn ngữ |

# DANH SÁCH BẢNG

[**Bảng 4. 1: Kết quả Rouge-Score sau khi train model** 34](#_Toc174221123)

# DANH SÁCH HÌNH VẼ

# MỞ ĐẦU

## Tính cấp thiết của đề tài

Thời đại số hóa thông tin mà chúng ta đang sống được đặc trưng bởi sự tăng trưởng nhanh chóng của lượng dữ liệu và thông tin được thu thập, lưu trữ và cung cấp dưới định dạng điện tử. Đa số dữ liệu doanh nghiệp được lưu trữ trong các tài liệu văn bản mà hầu hết là không có cấu trúc. Theo một nghiên cứu của Merrill Lynch và Gartner, 85% tổng số dữ liệu doanh nghiệp được thu thập và lưu trữ dưới dạng không có cấu trúc (McKnight, 2005). Cùng một nghiên cứu cũng cho biết rằng dữ liệu không có cấu trúc này đang tăng gấp đôi kích thước của nó mỗi 18 tháng. Bởi vì tri thức là quyền lực trong thế giới kinh doanh ngày nay, và tri thức được tạo ra từ dữ liệu và thông tin, các doanh nghiệp có khả năng tận dụng hiệu quả nguồn dữ liệu văn bản của họ sẽ có tri thức cần thiết để đưa ra quyết định tốt hơn, dẫn đến lợi thế cạnh tranh so với những doanh nghiệp kém phát triển. Đây là nơi mà nhu cầu về khai thác văn bản (Text Mining) phù hợp với bức tranh tổng thể của doanh nghiệp ngày nay.

Việc tóm tắt văn bản tiếng Việt nhằm giải quyết vấn đề tràn ngập thông tin trong thời đại hiện nay, sử dụng các công nghệ máy học và trí tuệ nhân tạo để tiết kiệm thời gian và tăng hiệu suất trong việc nắm bắt thông tin, đồng thời áp dụng ứng dụng rộng rãi trong các lĩnh vực khác nhau.

## Tổng quan về vấn đề nghiên cứu

### Khái quát ngắn gọn tổng quan về vấn đề nghiên cứu

Việc tóm tắt một đoạn văn là trình bày các điểm chính của nó một cách ngắn gọn. Công việc tự động tóm tắt văn bản bắt đầu hơn 40 năm trước. Sự phát triển của Internet đã thúc đẩy công việc này trong những năm gần đây, và các hệ thống tóm tắt bắt đầu được áp dụng trong các lĩnh vực như chăm sóc sức khỏe và thư viện số. Hiện nay, có nhiều chương trình tóm tắt văn bản thương mại trên thị trường. Các ví dụ bao gồm: ViT5 large, ViT5 base, BARTpho, mBART, mT5 và Transformer, ...

Đề tài **"Phát triển mô hình text mining dựa trên kỹ thuật Machine Learning cho tóm tắt văn bản tiếng Việt"** nghiên cứu về việc tóm tắt văn bản một cách tự động nhằm:

* **Tăng cường khả năng xử lý thông tin trong bối cảnh dữ liệu lớn**:
* Với sự bùng nổ của dữ liệu số trong thời đại hiện nay, việc xử lý và quản lý thông tin trở thành một thách thức lớn. Tóm tắt văn bản tự động là một công cụ mạnh mẽ giúp người dùng nhanh chóng nắm bắt được nội dung chính của một tài liệu hoặc tập hợp tài liệu lớn mà không cần phải đọc toàn bộ. Điều này đặc biệt quan trọng trong các lĩnh vực như báo chí, nghiên cứu khoa học, và quản lý tài liệu, nơi mà khối lượng thông tin cần xử lý là rất lớn.
* **Phát triển và ứng dụng của trí tuệ nhân tạo trong ngôn ngữ tự nhiên:**
* Ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một trong những lĩnh vực quan trọng của trí tuệ nhân tạo (AI), với nhiều ứng dụng trong thực tế như dịch máy, tìm kiếm thông tin, phân tích cảm xúc, và tóm tắt văn bản. Việc nghiên cứu và phát triển các mô hình tóm tắt văn bản tự động đóng góp vào sự phát triển chung của NLP, đồng thời mở ra các cơ hội ứng dụng mới trong nhiều ngành công nghiệp.
* **Ứng dụng thực tiễn trong nhiều lĩnh vực:**
* Tóm tắt văn bản tự động có thể được ứng dụng trong nhiều lĩnh vực khác nhau như giáo dục, kinh doanh, y tế, và công nghệ thông tin. Ví dụ, trong giáo dục, nó giúp học sinh, sinh viên nắm bắt nhanh nội dung tài liệu học tập. Trong kinh doanh, việc tóm tắt thông tin hỗ trợ quản lý thông tin từ báo cáo, nghiên cứu thị trường. Trong y tế, việc tóm tắt thông tin giúp các bác sĩ và nhà nghiên cứu nhanh chóng nắm bắt thông tin từ các tài liệu y khoa phức tạp.

### Khảo sát các công trình liên quan

*BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese (Nguyen Luong Tran, Duong Minh Le, Dat Quoc Nguyen)*, bài báo giới thiệu BARTpho với hai phiên bản, BARTphosyllable và BARTphoword, đây là những mô hình Sequence-to-Sequence đơn ngôn ngữ quy mô lớn công khai đầu tiên được tiền huấn luyện cho tiếng Việt. BARTpho sử dụng kiến trúc "large" và phương pháp tiền huấn luyện của mô hình tự giám sát Sequence-to-Sequence BART, do đó nó đặc biệt thích hợp cho các nhiệm vụ sinh văn bản trong lĩnh vực xử lý ngôn ngữ tự nhiên. Bài báo đã tiến hành các thí nghiệm để so sánh BARTpho của các tác giả với đối thủ mBART trên nhiệm vụ tóm tắt văn bản tiếng Việt và cho thấy rằng: trong cả đánh giá tự động và đánh giá từ con người, BARTpho vượt trội so với mBART, một mô hình cơ sở mạnh mẽ, và cải thiện trạng thái nghệ thuật hiện đại.

*ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation (Long Phan, Hieu Tran, Hieu Nguyen, Trieu H. Trinh),* các tác giả giới thiệu ViT5, một mô hình Transformer tiền huấn luyện Text-to-Text cho ngôn ngữ tiếng Việt. Sử dụng công thức tiền huấn luyện tự giám sát T5 trên các nguồn văn bản tiếng Việt lớn và chất lượng cao, nghiên cứu đã chỉ ra rằng ViT5 đạt được kết quả hàng đầu trên nhiệm vụ tóm tắt văn bản trên cả tập dữ liệu Wikilingua và Vietnews.

*Vietnamese doc summarization basic (Hoang Anh Pham),* tác giả đã giới thiệu một trong số những cách đơn giản nhất. Với việc áp dụng những phương pháp cơ bản nhất của học máy (Machine Learning) hay xử lý ngôn ngữ tự nhiên (Natural Language Processing). Tác giả thực hiện các bước: Preprocessing input document, Sentence tokenizer, Encode Sentences to vectors, Clustering, Build the summarization trên bộ dữ liệu neg.pkl – (5000 news and some papers in Vietnam) để xây dựng và hoàn thành chương trình tóm tắt văn bản (tiếng Việt).

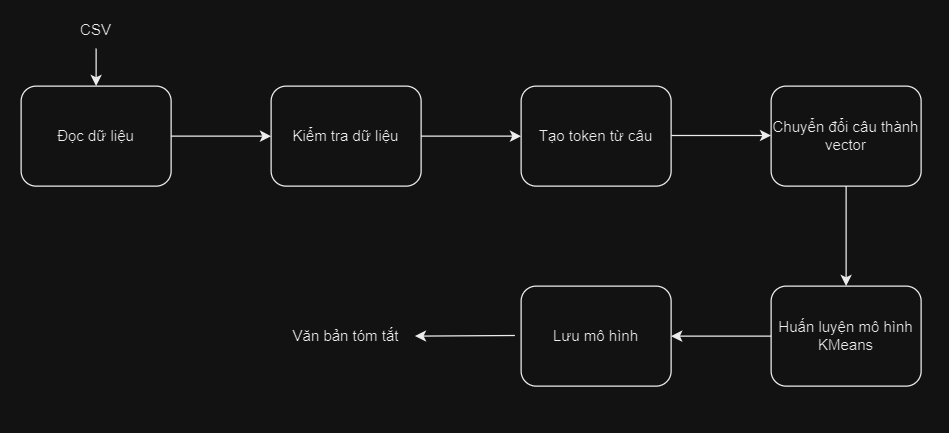
**Nhận xét***:* Thông qua khảo sát các công trình liên quan, theo sự hiểu biết tốt nhất của học viên, chủ đề “Phát triển mô hình text mining dựa trên kỹ thuật Machine Learning cho tóm tắt văn bản tiếng Việt” đang là một chủ đề còn rất nhiều vấn đề mở chưa được giải quyết. Xây dựng một mô hình tóm tắt văn bản tiếng Việt tối ưu và chính xác hơn có thể mang lại nhiều lợi ích, bao gồm cải thiện hiểu biết về thông tin, tăng cường kỹ năng viết và đọc, và hỗ trợ nhanh chóng trong việc tiếp cận nội dung phức tạp trên internet và trong văn học.

### Mục đích nghiên cứu

Mục tiêu chính của nghiên cứu này là phát triển một mô hình tóm tắt văn bản tự động hiệu quả cho tiếng Việt, đáp ứng các yêu cầu sau:

* Phát triển mô hình tóm tắt văn bản trích xuất dành riêng cho tiếng Việt:
* Thiết kế và xây dựng một mô hình tóm tắt văn bản trích xuất có khả năng lựa chọn và kết hợp các câu quan trọng nhất từ văn bản gốc để tạo ra bản tóm tắt ngắn gọn nhưng vẫn giữ được nội dung chính của văn bản.
* Đánh giá và so sánh với các mô hình hiện tại:
* So sánh mô hình được phát triển với các mô hình tóm tắt văn bản hiện có, để đánh giá khả năng áp dụng và hiệu quả của mô hình đối với tiếng Việt.

Mô hình học máy tóm tắt văn bản tiếng Việt dự kiến:



**Hình 1: Mô hình học máy tóm tắt văn bản tiếng Việt**

## Đối tượng và phạm vi nghiên cứu

### Đối tượng nghiên cứu

Các đối tượng nghiên cứu để phát triển một mô hình sử dụng Text Mining và machine learning cho tóm tắt văn bản tiếng Việt:

* Tiền xử lý văn bản: nltk và pyvi
* Chuyển đổi văn bản thành vector: FastText và TF-IDF
* Xử lý và tóm tắt văn bản: KMeans
* Đánh giá mô hình tóm tắt: ROUGE

### Phạm vi nghiên cứu

Bộ dữ liệu**: VNDS (A Vietnamese Dataset for Summarization)** (Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, Nguyen Xuan Hoai)

Bộ dữ liệu này đã được công bố vào tháng 12 năm 2019 và có mã DOI là 10.1109/NICS48868.2019.9023886. Được giới thiệu tại Hội nghị Thứ sáu về Khoa học Thông tin và Máy tính của NAFOSTED năm 2019 (NICS). Bộ dữ liệu đã được chính thức công bố và chia sẻ trong cộng đồng nghiên cứu.

VNDS là một bộ dữ liệu tiếng Việt được thiết kế đặc biệt cho nhiệm vụ tóm tắt văn bản. Đây là một nguồn tài nguyên quan trọng giúp nghiên cứu và phát triển các hệ thống tóm tắt văn bản tự động trong tiếng Việt. Bộ dữ liệu này cung cấp các tài liệu nguồn đa dạng, bao gồm nhiều chủ đề khác nhau, để phản ánh sự đa dạng của ngôn ngữ và văn hóa tiếng Việt.

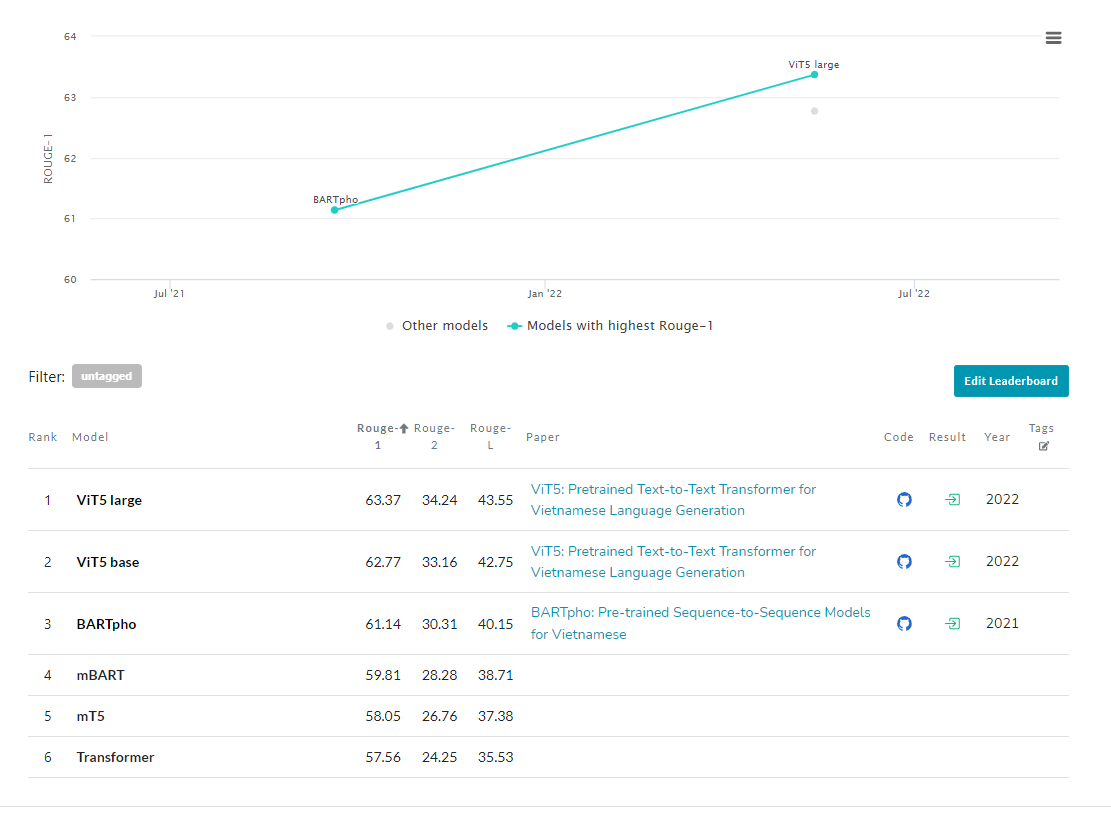
Các tài liệu trong VNDS được chú trọng để đảm bảo rằng chúng bao gồm thông tin chính và ý nghĩa tổng thể của văn bản nguồn. Bộ dữ liệu này cung cấp cơ hội cho các nhà nghiên cứu tiếng Việt nghiên cứu các phương pháp tóm tắt văn bản và đánh giá hiệu suất của chúng trên các tài liệu tiếng Việt thực tế.

Ngoài ra, VNDS không chỉ là một nguồn tài nguyên quan trọng cho việc phát triển và đánh giá các mô hình tóm tắt văn bản, mà còn giúp tạo ra một tiêu chuẩn trong lĩnh vực nghiên cứu tiếng Việt và thúc đẩy sự tiến bộ trong lĩnh vực này.

## Phương pháp nghiên cứu

Dựa vào các công trình nghiên cứu liên quan, học viên xây dựng mô hình tóm tắt văn bản tiếng Việt trên bộ dữ liệu **VNDS (VNDS: A Vietnamese Dataset for Summarization).** Học viên muốn tích hợp các kỹ thuật và phương pháp khác nhau trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và học máy, để đạt được mục tiêu tóm tắt văn bản. Điều này cho thấy một hướng tiếp cận đa dạng và linh hoạt để giải quyết vấn đề tóm tắt văn bản tiếng Việt. Cách tiếp cận này cho phép học tận dụng các công cụ và kỹ thuật phù hợp nhất với đặc điểm cụ thể của dữ liệu tiếng Việt.

Sau khi thực hiện xây dựng thành công mô hình, học viên thực hiện đánh giá so sánh mô hình đã được xây dựng với những nghiên cứu của các tác giả khác như: ViT5 large, ViT5 base, BARTpho, mBART, mT5, Transformer… trên cùng bộ dữ liệu VNDS (*như hình …*). Từ đó, đưa ra nhận xét, hướng phát triển để cải thiện mô hình.



**Hình 2: Nghiên cứu của các tác giả trên tập dữ liệu VNDS**

(Nguồn: VNDS: A Vietnamese Dataset for Summarization trên Kaggle.com)

## Bố cục đề án

Bên cạnh phần mở đầu, phần kết luận và phần tài liệu tham khảo, phần nội dung chính của đề án được chia thành 3 chương chính như sau:

Chương 1: Nghiên cứu tổng quan

Chương 2: Các thuật toán liên quan

Chương 3: Tổng kết và thực nghiệm chương trình

# CHƯƠNG 1: NGHIÊN CỨU TỔNG QUAN

## Tổng quan về Text Mining

### Text Mining là gì?

Khai thác văn bản (text mining), còn được gọi là khai thác dữ liệu văn bản, là quá trình chuyển đổi văn bản không cấu trúc thành định dạng có cấu trúc để xác định các mẫu ý nghĩa và thông tin mới. Chúng ta có thể sử dụng khai thác văn bản để phân tích các tập hợp lớn các tài liệu văn bản nhằm nắm bắt các khái niệm chính, xu hướng và mối quan hệ ẩn.

Bằng cách áp dụng các kỹ thuật phân tích nâng cao, chẳng hạn như Naïve Bayes, Support Vector Machines (SVM) và các thuật toán học sâu khác, các công ty có thể khám phá và phát hiện các mối quan hệ ẩn trong dữ liệu không cấu trúc của họ.

Văn bản là một trong những loại dữ liệu phổ biến nhất trong cơ sở dữ liệu. Tùy thuộc vào cơ sở dữ liệu, dữ liệu này có thể được tổ chức theo các định dạng sau:

* **Dữ liệu có cấu trúc**: Dữ liệu này được chuẩn hóa thành định dạng bảng với nhiều hàng và cột, giúp dễ lưu trữ và xử lý cho phân tích và các thuật toán học máy. Dữ liệu có cấu trúc có thể bao gồm các thông tin như tên, địa chỉ và số điện thoại.
* **Dữ liệu không cấu trúc**: Dữ liệu này không có định dạng dữ liệu được xác định trước. Nó có thể bao gồm văn bản từ các nguồn như mạng xã hội hoặc đánh giá sản phẩm, hoặc các định dạng phương tiện phong phú như video và tệp âm thanh.
* **Dữ liệu bán cấu trúc**: Như tên gọi, dữ liệu này là sự kết hợp giữa các định dạng dữ liệu có cấu trúc và không cấu trúc. Mặc dù nó có một số tổ chức, nhưng không đủ cấu trúc để đáp ứng yêu cầu của cơ sở dữ liệu quan hệ. Các ví dụ về dữ liệu bán cấu trúc bao gồm các tệp XML, JSON và HTML.

Vì khoảng 80% dữ liệu trên thế giới tồn tại dưới dạng không cấu trúc, khai thác văn bản là một thực hành vô cùng giá trị trong các tổ chức. Các công cụ khai thác văn bản và các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), như trích xuất thông tin, cho phép chúng ta chuyển đổi các tài liệu không cấu trúc thành định dạng có cấu trúc để phân tích và tạo ra những thông tin chất lượng cao. Điều này, lần lượt, cải thiện quy trình ra quyết định của các tổ chức, dẫn đến kết quả kinh doanh tốt hơn.

### Text mining và Text analytics

**Text Mining**

* **Định nghĩa**: Khai thác văn bản là quá trình chuyển đổi dữ liệu văn bản không cấu trúc thành định dạng có cấu trúc để phát hiện các mẫu, xu hướng và mối quan hệ tiềm ẩn. Quá trình này thường sử dụng các phương pháp xử lý ngôn ngữ tự nhiên (NLP) để phân tích nội dung văn bản và trích xuất thông tin có giá trị.
* **Mục tiêu**: Tìm kiếm và khám phá thông tin ẩn trong dữ liệu văn bản, như các chủ đề, mối quan hệ giữa các đối tượng, và các xu hướng tiềm năng.
* **Phương pháp**: Các kỹ thuật bao gồm phân tích ngữ nghĩa, trích xuất thông tin, phân loại văn bản, và phân cụm văn bản. Công cụ phổ biến bao gồm các thuật toán học máy và các phương pháp thống kê.

**Text analytics**

* **Định nghĩa:** Phân tích văn bản là quá trình phân tích các dữ liệu văn bản đã được chuẩn bị để rút ra các thông tin định lượng và chi tiết. Nó sử dụng các kỹ thuật phân tích và trực quan hóa dữ liệu để chuyển đổi thông tin văn bản thành những hiểu biết có thể hành động.
* **Mục tiêu:** Tạo ra các báo cáo và biểu đồ trực quan từ dữ liệu văn bản, cung cấp thông tin chi tiết và định lượng về các mẫu và xu hướng trong dữ liệu.
* **Phương pháp:** Sử dụng các công cụ phân tích thống kê, mô hình hóa dữ liệu, và trực quan hóa dữ liệu để hiểu và trình bày thông tin từ dữ liệu văn bản.

### Các kỹ thuật khai thác văn bản

Quá trình khai thác văn bản bao gồm nhiều hoạt động cho phép chúng ta suy luận thông tin từ dữ liệu văn bản không cấu trúc. Trước khi áp dụng các kỹ thuật khai thác văn bản khác nhau, chúng ta cần bắt đầu với bước tiền xử lý văn bản, là thực hành làm sạch và chuyển đổi dữ liệu văn bản thành định dạng có thể sử dụng. Đây là một khía cạnh cốt lõi của xử lý ngôn ngữ tự nhiên (NLP) và thường bao gồm các kỹ thuật như nhận diện ngôn ngữ, phân tách từ (tokenization), gán nhãn phần của lời (part-of-speech tagging), phân đoạn (chunking) và phân tích cú pháp (syntax parsing) để định dạng dữ liệu phù hợp cho phân tích. Khi hoàn thành bước tiền xử lý văn bản, chúng ta có thể áp dụng các thuật toán khai thác văn bản để rút ra thông tin từ dữ liệu. Một số kỹ thuật khai thác văn bản phổ biến bao gồm:

**Tìm kiếm thông tin (Information Retrieval)**

Tìm kiếm thông tin (IR) trả về thông tin hoặc tài liệu liên quan dựa trên một tập hợp các truy vấn hoặc cụm từ được xác định trước. Hệ thống IR sử dụng các thuật toán để theo dõi hành vi của người dùng và xác định dữ liệu liên quan. Tìm kiếm thông tin thường được sử dụng trong các hệ thống danh mục thư viện và các công cụ tìm kiếm phổ biến như Google. Một số nhiệm vụ phụ của IR bao gồm:

* **Phân tách từ (Tokenization)**: Là quá trình tách văn bản dài thành các câu và từ gọi là “tokens”. Những token này sau đó được sử dụng trong các mô hình như bag-of-words để phân cụm văn bản và các nhiệm vụ khớp tài liệu.
* **Rút gọn từ (Stemming)**: Là quá trình tách các tiền tố và hậu tố khỏi từ để tìm ra dạng gốc và ý nghĩa của từ. Kỹ thuật này cải thiện việc tìm kiếm thông tin bằng cách giảm kích thước của các tệp chỉ mục.

**Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)**

Xử lý ngôn ngữ tự nhiên, phát triển từ ngôn ngữ học tính toán, sử dụng các phương pháp từ nhiều lĩnh vực như khoa học máy tính, trí tuệ nhân tạo, ngôn ngữ học và khoa học dữ liệu, để giúp máy tính hiểu ngôn ngữ của con người cả dưới dạng viết và nói. Bằng cách phân tích cấu trúc câu và ngữ pháp, các nhiệm vụ phụ của NLP cho phép máy tính "đọc". Một số nhiệm vụ phụ phổ biến bao gồm:

* **Tóm tắt văn bản (Summarization):** Kỹ thuật này cung cấp một bản tóm tắt ngắn gọn của các văn bản dài để tạo ra một tóm tắt ngắn gọn, mạch lạc về các điểm chính của tài liệu.
* **Gán nhãn phần của lời (Part-of-Speech - PoS tagging):** Kỹ thuật này gán một thẻ cho từng token trong tài liệu dựa trên phần của lời của nó—tức là các danh từ, động từ, tính từ, v.v. Bước này cho phép phân tích ngữ nghĩa trên văn bản không cấu trúc.
* **Phân loại văn bản (Text Categorization):** Nhiệm vụ này, còn được gọi là phân loại văn bản, chịu trách nhiệm phân tích các tài liệu văn bản và phân loại chúng dựa trên các chủ đề hoặc danh mục đã định sẵn. Nhiệm vụ này đặc biệt hữu ích khi phân loại các từ đồng nghĩa và viết tắt.
* **Phân tích cảm xúc (Sentiment Analysis):** Nhiệm vụ này phát hiện cảm xúc tích cực hoặc tiêu cực từ các nguồn dữ liệu nội bộ hoặc bên ngoài, cho phép chúng ta theo dõi sự thay đổi trong thái độ của khách hàng theo thời gian. Nó thường được sử dụng để cung cấp thông tin về nhận thức của các thương hiệu, sản phẩm và dịch vụ. Những thông tin này có thể giúp các doanh nghiệp kết nối với khách hàng và cải thiện quy trình và trải nghiệm của người dùng.

**Trích xuất thông tin (Information Extraction - IE)**

Trích xuất thông tin (IE) tìm ra các phần dữ liệu liên quan khi tìm kiếm các tài liệu khác nhau. Nó cũng tập trung vào việc trích xuất thông tin có cấu trúc từ văn bản tự do và lưu trữ các thực thể, thuộc tính và thông tin mối quan hệ trong cơ sở dữ liệu. Một số nhiệm vụ phụ của trích xuất thông tin bao gồm:

* **Chọn đặc trưng (Feature Selection),** hay còn gọi là chọn thuộc tính, là quá trình chọn các đặc trưng quan trọng (kích thước) để đóng góp nhiều nhất vào kết quả của mô hình phân tích dự đoán.
* **Trích xuất đặc trưng (Feature Extraction)** là quá trình chọn một tập hợp con của các đặc trưng để cải thiện độ chính xác của nhiệm vụ phân loại. Đây là bước quan trọng trong việc giảm kích thước.
* **Nhận diện thực thể tên (Named-Entity Recognition - NER),** còn được gọi là nhận diện thực thể hoặc trích xuất thực thể, nhằm tìm và phân loại các thực thể cụ thể trong văn bản, chẳng hạn như tên hoặc địa điểm. Ví dụ, NER xác định “California” là một địa điểm và “Mary” là tên của một phụ nữ.

**Khai thác dữ liệu (Data Mining)**

Khai thác dữ liệu là quá trình xác định các mẫu và rút ra thông tin hữu ích từ các tập dữ liệu lớn. Thực hành này đánh giá cả dữ liệu có cấu trúc và không cấu trúc để xác định thông tin mới, và thường được sử dụng để phân tích hành vi của người tiêu dùng trong marketing và bán hàng. Khai thác văn bản về cơ bản là một lĩnh vực con của khai thác dữ liệu vì nó tập trung vào việc đưa cấu trúc cho dữ liệu không cấu trúc và phân tích nó để tạo ra các thông tin mới. Các kỹ thuật đã đề cập ở trên là các hình thức khai thác dữ liệu nhưng thuộc phạm vi phân tích dữ liệu văn bản.

### 1.1.4 Ứng dụng của khai thác văn bản

Text Mining có thể được áp dụng trong nhiều lĩnh vực khác nhau, từ nghiên cứu học thuật, phát hiện gian lận, phân tích tình cảm, đến phân tích thị trường và dự báo xu hướng. Điều này làm cho nó trở thành một công cụ đa năng và có giá trị cao:

**Information Retrieval (IR)**: là việc liên kết và truy xuất thông tin từ một số lượng lớn tài liệu dựa trên văn bản. IR khác với hệ thống cơ sở dữ liệu ở chỗ nó không gặp phải một số vấn đề như kiểm soát đồng thời, khôi phục, quản lý giao dịch và cập nhật. Ngược lại, IR phải xử lý các tài liệu không có cấu trúc, tìm kiếm ước lượng dựa trên từ khóa và khái niệm liên quan. Do lượng thông tin văn bản khổng lồ, IR đã được áp dụng rộng rãi trong các hệ thống như thư viện trực tuyến, hệ thống quản lý tài liệu trực tuyến và các công cụ tìm kiếm trên web.

**Categorization**: Phân loại là quá trình xác định các chủ đề chính của một tài liệu bằng cách đưa tài liệu vào một tập hợp các chủ đề đã được xác định trước. Khi phân loại một tài liệu, chương trình máy tính thường xử lý tài liệu như một "túi từ" ("bag of words"). Nó không cố gắng xử lý thông tin thực sự như trích xuất thông tin. Thay vào đó, phân loại chỉ đếm số lần xuất hiện của các từ và từ đó, từ đó nhận diện các chủ đề chính mà tài liệu bao gồm. Phân loại thường phụ thuộc vào một từ điển mà các chủ đề đã được xác định trước, và các mối quan hệ được xác định bằng cách tìm kiếm các thuật ngữ lớn, thuật ngữ hẹp hơn, từ đồng nghĩa và các thuật ngữ liên quan.

**Natural Language Processing**: là một lĩnh vực nghiên cứu và ứng dụng nghiên cứu cách thức mà máy tính có thể được sử dụng để hiểu và xử lý văn bản ngôn ngữ tự nhiên. Các nhà nghiên cứu NLP nhằm thu thập kiến ​​thức về cách con người hiểu và sử dụng ngôn ngữ để phát triển các công cụ và kỹ thuật phù hợp, giúp hệ thống máy tính hiểu và xử lý các ngôn ngữ tự nhiên để thực hiện các nhiệm vụ mong muốn.

## Tổng quan về NLP

Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực nghiên cứu và ứng dụng tìm hiểu cách máy tính có thể được sử dụng để hiểu và xử lý văn bản ngôn ngữ tự nhiên. Các nhà nghiên cứu NLP nhắm tới việc thu thập kiến thức về cách con người hiểu và sử dụng ngôn ngữ để phát triển các công cụ và kỹ thuật phù hợp, giúp các hệ thống máy tính có thể hiểu và xử lý các ngôn ngữ tự nhiên để thực hiện các nhiệm vụ mong muốn.

Các kiến thức cơ bản của NLP nằm trong nhiều lĩnh vực như khoa học máy tính và thông tin, ngôn ngữ học, toán học, kỹ thuật điện và điện tử, trí tuệ nhân tạo và robot, tâm lý học, v.v. Các ứng dụng của NLP bao gồm nhiều lĩnh vực nghiên cứu, chẳng hạn như dịch máy, xử lý và tóm tắt văn bản ngôn ngữ tự nhiên, giao diện người dùng, truy xuất thông tin đa ngôn ngữ và xuyên ngôn ngữ (CLIR), nhận dạng giọng nói, trí tuệ nhân tạo và hệ thống chuyên gia, v.v.

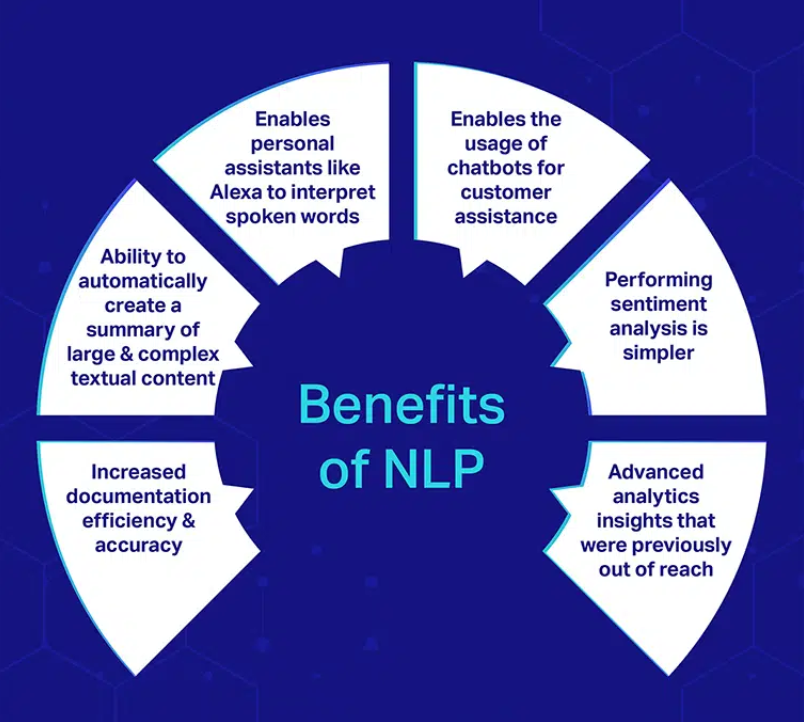
Một số ứng dụng text mining ban đầu đã sử dụng mô hình đơn giản gọi là bag-of-words khi tạo cấu trúc cho bộ sưu tập các tài liệu văn bản để phân loại chúng vào hai hoặc nhiều lớp đã xác định trước hoặc để phân nhóm chúng vào các nhóm tự nhiên. Trong mô hình bag-of-words, văn bản, như câu, đoạn văn hoặc tài liệu hoàn chỉnh, được biểu diễn như một tập hợp các từ, không quan tâm đến ngữ pháp hoặc thứ tự xuất hiện của các từ. Mô hình bag-of-words vẫn được sử dụng trong một số công cụ phân loại tài liệu đơn giản. Ví dụ, trong lọc thư rác, một thư điện tử có thể được mô hình hóa như một tập hợp các từ không có thứ tự (bag-of-words) và được so sánh với hai túi từ đã xác định trước. Một túi chứa các từ tìm thấy trong tin nhắn rác và túi kia chứa các từ tìm thấy trong thư điện tử hợp pháp. Mức độ khớp giữa bag-of-words của một email cụ thể và hai túi từ chứa các từ mô tả xác định liệu email đó là thư rác hay hợp pháp.

Tuy nhiên, con người không sử dụng từ ngữ mà không có thứ tự hay cấu trúc. Chúng ta sử dụng từ trong các câu, có cấu trúc ngữ nghĩa cũng như cú pháp. Vì vậy, các kỹ thuật tự động (như text mining) cần tìm cách vượt qua cách diễn giải bag-of-words và tích hợp ngày càng nhiều cấu trúc ngữ nghĩa vào hoạt động của chúng. Xu hướng hiện nay trong text mining là hướng tới việc bao gồm nhiều tính năng nâng cao có thể đạt được bằng cách sử dụng xử lý ngôn ngữ tự nhiên (NLP).

Natural language processing (NLP) là một thành phần quan trọng của text mining và là một nhánh con của trí tuệ nhân tạo và ngôn ngữ học tính toán. Nó nghiên cứu vấn đề "hiểu" ngôn ngữ tự nhiên của con người, với mục tiêu chuyển đổi các mô tả của ngôn ngữ con người (như các tài liệu văn bản) thành các biểu diễn chính thức hơn (dưới dạng dữ liệu số và ký hiệu) dễ dàng cho các chương trình máy tính thao tác. Mục tiêu của NLP là vượt ra khỏi việc thao tác văn bản dựa trên cú pháp (thường được gọi là "word counting") để đạt đến một sự hiểu biết và xử lý thực sự của ngôn ngữ tự nhiên, xem xét cả các ràng buộc ngữ pháp và ngữ nghĩa cũng như ngữ cảnh.

### Những lợi ích và thách thức liên quan đến việc triển khai NLP

**Lợi ích của NLP:**

****

**Hình 2. 1: Lợi ích của NLP**

(Nguồn: What is NLP? How it Works, Benefits, Challenges, Examples - Internet)

* **Tăng hiệu quả và độ chính xác của việc lập tài liệu: Một tài liệu được tạo bởi NLP có thể tóm tắt chính xác bất kỳ văn bản gốc nào mà con người không thể tự động tạo ra. Ngoài ra, nó có thể thực hiện các nhiệm vụ lặp đi lặp lại như phân tích các khối dữ liệu lớn để cải thiện hiệu quả của con người.**
* **Khả năng tự động tạo bản tóm tắt cho nội dung văn bản lớn và phức tạp:** Xử lý ngôn ngữ tự nhiên có thể được sử dụng cho các nhiệm vụ khai thác văn bản đơn giản như trích xuất thông tin từ tài liệu, phân tích cảm xúc, hoặc xác định các thực thể được đặt tên. Nó cũng có thể được sử dụng cho các nhiệm vụ phức tạp hơn, chẳng hạn như hiểu hành vi và cảm xúc của con người.
* **Cho phép các trợ lý cá nhân như Alexa hiểu được từ ngữ nói: NLP hữu ích cho các trợ lý cá nhân như Alexa, cho phép trợ lý ảo hiểu các lệnh bằng lời nói. Nó cũng giúp tìm kiếm nhanh chóng thông tin liên quan từ các cơ sở dữ liệu chứa hàng triệu tài liệu trong vài giây.**
* **Cho phép sử dụng chatbot để hỗ trợ khách hàng: NLP có thể được sử dụng trong các chatbot và các chương trình máy tính sử dụng trí tuệ nhân tạo để giao tiếp với con người qua văn bản hoặc giọng nói. Chatbot sử dụng NLP để hiểu những gì người dùng đang gõ và phản hồi một cách thích hợp. Chúng cũng cho phép một tổ chức cung cấp hỗ trợ khách hàng 24/7 trên nhiều kênh.**
* **Đơn giản hóa việc phân tích cảm xúc:** Phân tích cảm xúc là một quá trình phân tích một tập hợp các tài liệu (chẳng hạn như đánh giá hoặc tweet) liên quan đến thái độ hoặc trạng thái cảm xúc của chúng (ví dụ: vui mừng, tức giận). Phân tích cảm xúc có thể được sử dụng để phân loại và phân nhóm các bài đăng trên mạng xã hội hoặc các văn bản khác thành các danh mục như tích cực, tiêu cực hoặc trung lập.
* **Cung cấp thông tin phân tích nâng cao mà trước đây khó tiếp cận:** Sự phát triển gần đây của các cảm biến và thiết bị kết nối Internet đã dẫn đến sự bùng nổ về khối lượng và đa dạng của dữ liệu được tạo ra. Kết quả là nhiều tổ chức tận dụng NLP để hiểu dữ liệu của họ nhằm đưa ra các quyết định kinh doanh tốt hơn.

**Những thách thức với NLP**

****

**Hình 2. 2: Những thách thức với NLP**

(Nguồn: What is NLP? How it Works, Benefits, Challenges, Examples - Internet)

* **Lỗi chính tả: Ngôn ngữ tự nhiên đầy những lỗi chính tả, lỗi đánh máy, và không nhất quán trong phong cách. Ví dụ, từ "process" có thể được viết là "process" hoặc "processing." Vấn đề trở nên phức tạp hơn khi thêm các dấu hoặc các ký tự không có trong từ điển.**
* **Khác biệt ngôn ngữ: Một người nói tiếng Anh có thể nói, "I’m going to work tomorrow morning," trong khi một người nói tiếng Ý sẽ nói, "Domani Mattina vado al lavoro." Mặc dù hai câu này có nghĩa giống nhau, NLP sẽ không hiểu câu tiếng Ý trừ khi dịch nó sang tiếng Anh trước.**
* **Thiên vị bẩm sinh: Ngôn ngữ tự nhiên dựa trên logic và bộ dữ liệu của con người. Trong một số tình huống, hệ thống NLP có thể mang theo những thiên vị của người lập trình hoặc bộ dữ liệu mà chúng sử dụng. Nó cũng có thể đôi khi hiểu sai ngữ cảnh do những thiên vị bẩm sinh, dẫn đến kết quả không chính xác.**
* **Từ có nhiều nghĩa: NLP dựa trên giả định rằng ngôn ngữ là chính xác và không mơ hồ. Trong thực tế, ngôn ngữ không chính xác cũng không không mơ hồ. Nhiều từ có nhiều nghĩa và có thể được sử dụng theo nhiều cách khác nhau. Ví dụ, khi chúng ta nói "bark," nó có thể là tiếng sủa của chó hoặc vỏ cây.**
* **Sự không chắc chắn và kết quả dương tính giả: Kết quả dương tính giả xảy ra khi NLP phát hiện một thuật ngữ lẽ ra có thể hiểu được nhưng lại không thể phản hồi chính xác. Mục tiêu là tạo ra một hệ thống NLP có thể xác định được các giới hạn của nó và làm rõ sự nhầm lẫn bằng cách sử dụng câu hỏi hoặc gợi ý.**
* **Dữ liệu đào tạo: Một trong những thách thức lớn nhất với xử lý ngôn ngữ tự nhiên là dữ liệu đào tạo không chính xác. Càng có nhiều dữ liệu đào tạo, kết quả sẽ càng tốt. Nếu chúng ta cung cấp cho hệ thống dữ liệu không chính xác hoặc thiên vị, nó sẽ hoặc học những điều sai hoặc học không hiệu quả.**

### Những công việc phổ biến áp dụng NLP



**Hình 2. 3: Những công việc phổ biến áp dụng NLP**

(Nguồn: What is NLP? How it Works, Benefits, Challenges, Examples - Internet)

**Nhận diện giọng nói:** Điều này liên quan đến việc chuyển đổi dữ liệu giọng nói hoặc âm thanh thành văn bản. Quá trình này rất quan trọng đối với bất kỳ ứng dụng NLP nào có tính năng lệnh giọng nói. Nhận diện giọng nói giải quyết sự đa dạng trong phát âm, phương ngữ, tốc độ, âm lượng, giọng điệu và các yếu tố khác để giải mã thông điệp dự định.

**Gán nhãn từ loại:** Tương tự như cách chúng ta được dạy các nguyên tắc cơ bản về ngữ pháp ở trường, điều này dạy máy móc xác định các thành phần từ loại trong câu như danh từ, động từ, tính từ và nhiều hơn nữa. Nó cũng dạy hệ thống hiểu khi nào một từ được sử dụng như một động từ và khi từ đó được sử dụng như một danh từ.

**Phân biệt nghĩa của từ:** Đây là một quá trình quan trọng chịu trách nhiệm về việc hiểu đúng nghĩa của một câu. Sử dụng phân tích ngữ nghĩa trong nhiệm vụ này cho phép máy hiểu liệu một người đã nói "Mọi thứ đang diễn ra tốt đẹp" như một lời châm biếm khi đang trải qua khủng hoảng.

**Nhận dạng thực thể có tên:** Khi có nhiều trường hợp danh từ như tên, địa điểm, quốc gia, và nhiều hơn nữa, một quá trình gọi là Nhận dạng thực thể có tên được triển khai. Quá trình này xác định và phân loại các thực thể trong một thông điệp hoặc lệnh và gia tăng giá trị cho việc hiểu của máy.

**Giải quyết đồng tham chiếu:** Con người thường rất sáng tạo khi giao tiếp, vì vậy có rất nhiều ẩn dụ, so sánh, cụm động từ và thành ngữ. Tất cả các sự mơ hồ phát sinh từ những điều này được làm rõ bằng nhiệm vụ Giải quyết đồng tham chiếu, giúp máy học rằng nó không thực sự "mưa mèo và chó" mà là đề cập đến cường độ của mưa.

**Tạo ngôn ngữ tự nhiên:** Nhiệm vụ này liên quan đến việc tạo ra văn bản giống như con người từ dữ liệu. Văn bản này có thể được tùy chỉnh theo tiếng lóng, ngôn ngữ khu vực, và nhiều hơn nữa.

# CHƯƠNG 2: CÁC THUẬT TOÁN LIÊN QUAN

## 2.1 Các bước xử lý Text Mining

### 2.1.1 Thu thập dữ liệu

* Thu thập dữ liệu từ nhiều nguồn khác nhau như văn bản, trang web, cơ sở dữ liệu, mạng xã hội.

### 2.1.2 Tiền xử lý dữ liệu

* Loại bỏ nhiễu: Loại bỏ các thành phần không cần thiết như HTML tags, ký tự đặc biệt, số, hoặc các từ dừng (stopwords).
* Chuẩn hóa văn bản: Chuyển đổi văn bản về dạng thống nhất, như viết thường tất cả chữ cái, chuyển đổi số thành từ.
* Tách từ: Chia văn bản thành các từ hoặc cụm từ có ý nghĩa.
* Gán nhãn từ loại: Xác định loại từ của từng từ trong văn bản (danh từ, động từ, tính từ).

### 2.1.3 Biểu diễn văn bản

* **Bag of Words** (BoW): Biểu diễn văn bản dưới dạng vector số, dựa trên tần suất xuất hiện của các từ.
* **TF-IDF** (Term Frequency-Inverse Document Frequency): Phương pháp đánh trọng số cho mỗi từ trong văn bản dựa trên tần suất xuất hiện và tần suất xuất hiện trong toàn bộ tập dữ liệu.
* **Word Embeddings:** Sử dụng các mô hình như Word2Vec, GloVe hoặc BERT để biểu diễn từ dưới dạng vector với ngữ nghĩa liên quan.

### 2.1.4 Phân tích và khai phá dữ liệu

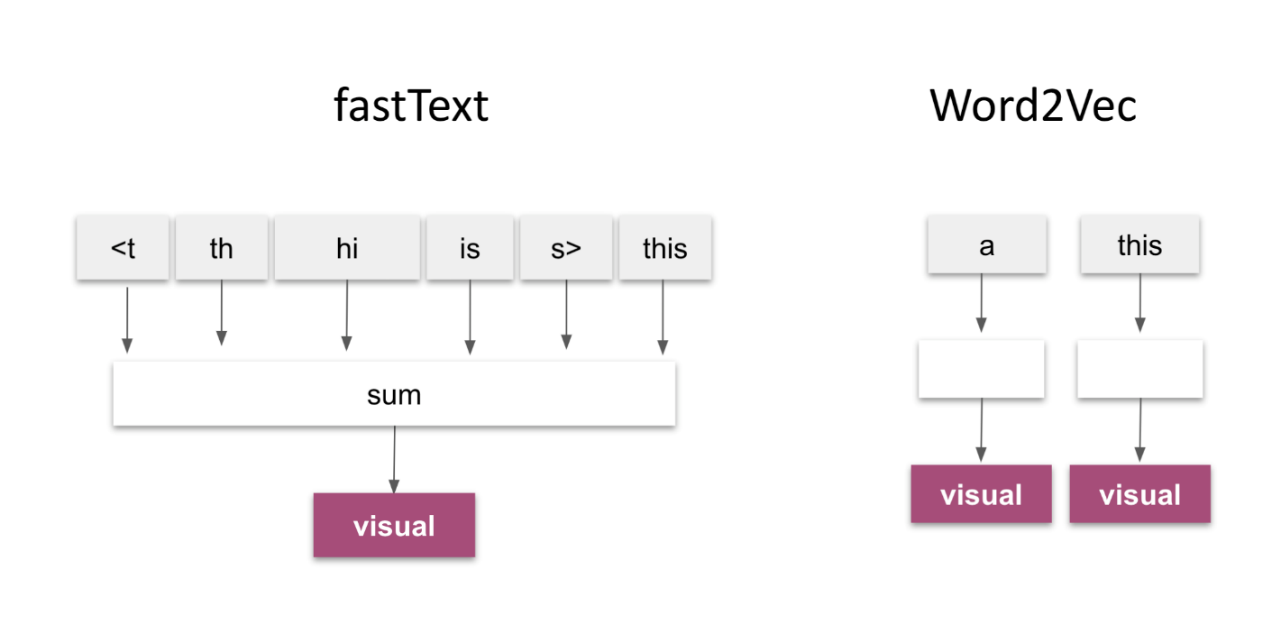
* Phân cụm văn bản (Text Clustering): Nhóm các văn bản tương tự nhau thành các cụm.
* Phân loại văn bản (Text Classification): Xác định nhãn hoặc chủ đề cho các văn bản dựa trên nội dung.
* Trích xuất thực thể có tên (Named Entity Recognition - NER): Xác định và phân loại các thực thể như tên người, địa điểm, tổ chức, v.v.
* Phân tích cảm xúc (Sentiment Analysis): Xác định cảm xúc hoặc thái độ của người viết trong văn bản.

### 2.1.5 Đánh giá kết quả

* Sử dụng các chỉ số đánh giá như độ chính xác, độ phủ, F1-score, ROUGE, để đánh giá hiệu quả của mô hình.

## 2.2 Thuật toán và độ đo

### 2.2.1 FastText thay vì Word2Vec



**Hình 3. 1: FastText & Word2Vec**

(Nguồn: FastText vs. Word2vec: A Quick Comparison - Internet)

**Word2Vec**

* **Ưu điểm:**

Hiệu suất cao: Word2Vec có thể xử lý các tập dữ liệu lớn nhanh chóng và hiệu quả.

Biểu diễn ngữ nghĩa mạnh mẽ: Dễ dàng nắm bắt các mối quan hệ ngữ nghĩa giữa các từ thông dụng.

* **Nhược điểm:**

Không xử lý tốt các từ hiếm và từ mới: Do không xem xét các thành phần con của từ, Word2Vec gặp khó khăn với các từ ít gặp hoặc từ mới.

Hạn chế trong ngôn ngữ có cấu trúc phức tạp: Tiếng Việt có nhiều từ ghép và từ có dấu, khiến cho việc biểu diễn từ ngữ trở nên phức tạp.

**FastText**

* **Ưu điểm:**

Xử lý tốt các từ hiếm và từ mới: Nhờ vào việc sử dụng các subword (n-grams), fastText có thể nắm bắt được các từ ngữ mới và hiếm gặp.

Hiệu quả trong ngôn ngữ phức tạp: Có thể phân tích và biểu diễn tốt các từ ngữ có dấu và từ ghép trong tiếng Việt.

Biểu diễn ngữ nghĩa chính xác hơn: Các vector từ ngữ trong fastText thường có độ chính xác cao hơn trong nhiều tác vụ xử lý ngôn ngữ tự nhiên.

* **Nhược điểm:**

Tốc độ huấn luyện chậm hơn: Do phải xử lý các subword, fastText có tốc độ huấn luyện chậm hơn so với Word2Vec.

Độ phức tạp tính toán cao hơn: Việc tính toán các vector subword phức tạp hơn, yêu cầu tài nguyên tính toán nhiều hơn.

* Dựa trên các ưu điểm của fastText trong việc xử lý ngôn ngữ tiếng Việt, bao gồm khả năng nắm bắt các từ ngữ hiếm và mới, cũng như xử lý tốt các đặc điểm ngôn ngữ phức tạp như từ ghép và dấu, lựa chọn fastText là hợp lý và tối ưu cho việc biểu diễn từ ngữ trong các tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt.

### 2.2.2 Kmean Clustering.

Trong thuật toán k-Means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm. Chúng ta sẽ dựa vào khoảng cách từ mỗi quan sát tới các tâm để xác định nhãn cho chúng trùng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượt cập nhật tâm chạm ngưỡng.

Các bước của thuật toán k-Means được tóm tắt như sau:

1. Khởi tạo ngẫu nhiên 𝑘 tâm cụm 𝜇1,𝜇2,…,𝜇𝑘.
2. Lặp lại quá trình cập nhật tâm cụm cho tới khi dừng:
   1. Xác định nhãn cho từng điểm dữ liệu 𝑐𝑖 dựa vào khoảng cách tới từng tâm cụm:

(Nguồn: Deep AI KhanhBlog - Internet)

* 1. Tính toán lại tâm cho từng cụm theo trung bình của toàn bộ các điểm dữ liệu trong một cụm:

(Nguồn: Deep AI KhanhBlog - Internet)

**So sánh với các phương pháp khác:**

Latent Semantic Analysis (LSA): Phương pháp này phân tích các mối quan hệ giữa các từ trong văn bản và có thể tạo ra tóm tắt chất lượng cao. Tuy nhiên, LSA có thể phức tạp hơn và đòi hỏi nhiều tài nguyên tính toán hơn so với K-means.

LexRank và TextRank: Đây là các phương pháp dựa trên đồ thị để tóm tắt văn bản, rất hiệu quả và không cần phải xác định số cụm trước. Tuy nhiên, các phương pháp này có thể trở nên phức tạp và khó triển khai hơn so với K-means.

Deep Learning: Các mô hình tóm tắt văn bản dựa trên deep learning như BERTSUM hoặc T5 có thể tạo ra tóm tắt chất lượng rất cao nhưng đòi hỏi nhiều tài nguyên tính toán và dữ liệu đào tạo lớn. K-means là một giải pháp đơn giản và nhẹ nhàng hơn trong nhiều trường hợp.

* Sử dụng K-means trong tóm tắt văn bản tiếng Việt là một lựa chọn hợp lý vì một phương pháp phân cụm đơn giản, hiệu quả và có thể mở rộng. Nó đảm bảo bản tóm tắt bao quát được nhiều khía cạnh khác nhau của văn bản gốc và có thể được điều chỉnh linh hoạt cho nhiều loại văn bản khác nhau.

### 2.2.3 Distance Measurement.

Kỹ thuật: ***pairwise\_distances\_argmin\_min*** để đo khoảng cách giữa các vector câu và tâm cụm.

KMeans sử dụng khoảng cách Euclidean để đo khoảng cách giữa các điểm và các tâm cụm:

(Nguồn: Deep AI KhanhBlog - Internet)

### 2.2.4 Term Frequency-Inverse Document Frequency.

Term Frequency-Inverse Document Frequency (TF-IDF) là một phương pháp phổ biến để biểu diễn các từ ngữ dưới dạng vector trong tài liệu văn bản. Phương pháp này kết hợp hai yếu tố chính: tần suất xuất hiện của một từ trong một tài liệu (Term Frequency - TF) và tầm quan trọng của từ đó trên toàn bộ tập hợp các tài liệu (Inverse Document Frequency - IDF). Khi sử dụng TF-IDF để tạo vector cho câu, ta có thể xác định tầm quan trọng của từng từ trong câu đó dựa trên tần suất xuất hiện và mức độ hiếm gặp của từ trong toàn bộ tập dữ liệu.

* **Term Frequency (TF):** Đo lường tần suất xuất hiện của một từ trong một tài liệu.

(Nguồn: TF-IDF ( term frequency – inverse document frequency)- Viblo - Internet)

* **Inverse Document Frequency (IDF):** Đo lường tầm quan trọng của một từ trên toàn bộ tập hợp các tài liệu.

(Nguồn: TF-IDF ( term frequency – inverse document frequency)- Viblo - Internet)

* **TF-IDF:** Kết hợp TF và IDF để đo lường tầm quan trọng của một từ trong một tài liệu cụ thể.

(Nguồn: TF-IDF ( term frequency – inverse document frequency)- Viblo - Internet)

* TF-IDF là một phương pháp mạnh mẽ để biểu diễn từ ngữ trong văn bản, cho phép xác định tầm quan trọng của từ ngữ dựa trên tần suất xuất hiện và mức độ hiếm gặp trong toàn bộ tập dữ liệu. Sử dụng TF-IDF để tạo vector cho câu giúp nắm bắt được thông tin ngữ nghĩa và tầm quan trọng của từ ngữ, hữu ích trong nhiều tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, tìm kiếm thông tin và phân cụm văn bản.

### 2.2.5 ROUGE Score.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một bộ các thước đo phổ biến để đánh giá chất lượng của hệ thống tóm tắt văn bản tự động. ROUGE đo lường sự trùng khớp giữa bản tóm tắt do máy tạo ra và bản tóm tắt tham chiếu do con người viết. ROUGE bao gồm các biến thể như ROUGE-N, ROUGE-L, và ROUGE-S. Dưới đây là cách tính toán và sử dụng các biến thể này:

**ROUGE-N:**

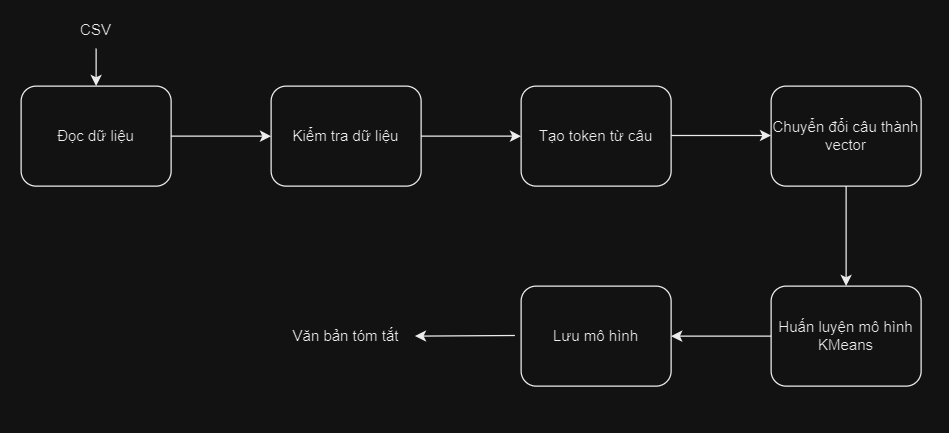
(Nguồn: 2 phút để hiểu cách tính ROUGE metric – Viblo - Internet)

**ROUGE-1:** Đo lường sự trùng khớp của các unigram (từ đơn lẻ) giữa bản tóm tắt máy và bản tóm tắt tham chiếu.

**ROUGE-2**: Đo lường sự trùng khớp của các bigram (cặp từ liên tiếp) giữa bản tóm tắt máy và bản tóm tắt tham chiếu.

* Việc sử dụng các thước đo ROUGE trong đánh giá model tóm tắt văn bản có những ưu điểm sau:
* Khả năng đánh giá tự động: ROUGE cho phép đánh giá chất lượng của các hệ thống tóm tắt một cách tự động và nhanh chóng, giảm bớt sự phụ thuộc vào đánh giá thủ công.
* Tính khách quan: Các thước đo ROUGE dựa trên sự trùng khớp của các n-gram hoặc chuỗi con chung, mang lại sự khách quan trong đánh giá.
* Khả năng so sánh: ROUGE cho phép so sánh hiệu suất giữa các hệ thống tóm tắt khác nhau một cách dễ dàng.
* Độc lập ngôn ngữ: ROUGE có thể được áp dụng cho nhiều ngôn ngữ khác nhau, bao gồm cả tiếng Việt.

## 2.3 Giải pháp đề xuất tóm tắt văn bản



* Sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để biểu diễn các câu dưới dạng vector, sau đó phân cụm các câu để chọn ra những câu quan trọng nhất để tạo bản tóm tắt. Sự kết hợp giữa TF-IDF và FastText giúp giữ lại cả tần suất và ngữ nghĩa của các từ trong quá trình tóm tắt.
* **Sử dụng KMeans Clustering**: Sau khi chuyển đổi các câu thành vector, thực hiện phân cụm các câu bằng KMeans để chọn các câu đại diện cho bản tóm tắt.
* KMeans giúp phân nhóm các câu tương tự lại với nhau và chọn câu đại diện từ mỗi nhóm. Điều này có thể tạo ra tóm tắt cô đọng và chính xác hơn so với các phương pháp chỉ chọn câu đơn lẻ.
* **Kết hợp TF-IDF và FastText**: Trong việc tính vector trung bình có trọng số cho toàn bộ câu.
* Với mỗi từ trong câu, nhân vector FastText của từ đó với trọng số TF-IDF của nó.
* Tổng hợp các vector từ đã nhân với TF-IDF bằng cách cộng chúng lại.
* Cuối cùng, chia tổng vector này cho tổng trọng số TF-IDF của các từ trong câu để tạo ra vector trung bình có trọng số cho cả câu.
* Vector cuối cùng của câu là một vector có trọng số, kết hợp thông tin ngữ nghĩa (từ FastText) và mức độ quan trọng (từ TF-IDF) của các từ trong câu.
* Kết hợp trọng số TF-IDF để điều chỉnh độ quan trọng của từ và mô hình FastText để giữ ngữ nghĩa của từ. Sử dụng cả TF-IDF và FastText giúp mô hình không chỉ cân nhắc trọng số của từ mà còn giữ lại ngữ nghĩa từ, tạo ra các vector câu giàu thông tin hơn

# CHƯƠNG 3: TỔNG KẾT VÀ THỰC NGHIỆM CHƯƠNG TRÌNH

## 3.1 Thực nghiệm chương trình.

### 3.1.1 Đọc dữ liệu

* Tải dữ liệu huấn luyện và kiểm tra từ các tệp CSV (Bộ dữ liệu**: VNDS: A Vietnamese Dataset for Summarization)** để chuẩn bị cho các bước xử lý tiếp theo.

### 3.1.2 Kiểm tra dữ liệu

* Kiểm tra kích thước (số dòng và cột) của dữ liệu huấn luyện và kiểm tra để đảm bảo dữ liệu được tải lên đúng cách.

### 3.1.3 Tạo token từ câu

* **Chuẩn hóa văn bản:** Chuyển văn bản về chữ thường, loại bỏ các ký tự không cần thiết (khoảng trắng thừa, ký tự đặc biệt, dấu câu).
* **Tách câu:** Sử dụng nltk.sent\_tokenize để tách văn bản thành các câu.
* **Xử lý từng câu:** Loại bỏ các ký tự không thuộc từ (sử dụng biểu thức chính quy re.sub) và chuẩn hóa khoảng trắng trong câu.
* Chuẩn hóa và tách câu từ văn bản gốc để chuẩn bị cho việc chuyển đổi câu thành vector.

### 3.1.4 Chuyển đổi câu thành vector

* **Tạo TF-IDF vectorizer:** Sử dụng TfidfVectorizer để tính trọng số TF-IDF cho từng từ trong câu, giúp xác định tầm quan trọng của từ đó trong ngữ cảnh của toàn bộ văn bản.
* **Chuyển đổi từ thành vector bằng FastText:** Mỗi từ sau khi được token hóa sẽ được chuyển thành một vector số học (chuỗi các số) bằng mô hình FastText.
* **Nhân vector từ với trọng số TF-IDF:** Mỗi vector từ được nhân với trọng số TF-IDF tương ứng của từ đó. Điều này giúp điều chỉnh độ quan trọng của từ trong câu. Những từ quan trọng hơn sẽ có trọng số TF-IDF cao hơn và sẽ có ảnh hưởng lớn hơn đến vector cuối cùng của câu.
* **Chuẩn hóa vector câu:** Sau khi tính toán xong, vector của câu được chuẩn hóa bằng cách chia cho tổng trọng số TF-IDF của các từ. Điều này đảm bảo rằng vector cuối cùng không quá phụ thuộc vào độ dài của câu, mà tập trung vào ý nghĩa tổng quát của câu.
* Biểu diễn các câu thành các vector số, trong đó mỗi vector phản ánh ý nghĩa ngữ nghĩa của câu dựa trên mô hình từ điển (FastText) và trọng số TF-IDF.

### 3.1.5 Huấn luyện mô hình Kmeans

* **Phân cụm các câu:** Sử dụng KMeans để phân cụm các câu thành 2, 3, hoặc 4 cụm, tùy thuộc vào độ phù hợp của từng số cụm.
* **Chọn câu tiêu biểu:** Tính trung bình các chỉ số của cụm và chọn câu gần nhất với tâm cụm để làm câu đại diện cho cụm đó.
* **Tạo tóm tắt:** Kết hợp các câu đại diện để tạo thành đoạn tóm tắt cho đoạn văn.
* Phân cụm các câu trong đoạn văn thành các nhóm tương tự để tóm tắt đoạn văn.

### 3.1.6 Lưu mô hình

* Lưu mô hình KMeans và các nhãn

## 3.2 Kết luận và hướng phát triển.

* Kết quả đánh giá mô hình:

|  |  |  |
| --- | --- | --- |
|  | Data train | Data test |
| precision score | Rouge\_1: 32.03590517623662  Rouge\_2: 16.295291882382653  Rouge\_L: 18.07967891795754 | Rouge\_1: 30.196701317606077  Rouge\_2: 16.777715767421411  Rouge\_L: 17.255594820293226 |
| recall score | Rouge\_1: 67.75772334994721  Rouge\_2: 24.323272066192915  Rouge\_L: 39.690100330098154 | Rouge\_1: 73.82706603196205  Rouge\_2: 28.058460515808502  Rouge\_L: 44.04807771533466 |
| fmeasure score | Rouge\_1: 41.41591770123297  Rouge\_2: 18.346944529568876  Rouge\_L: 23.626354486977633 | Rouge\_1: 41.4590548102516  Rouge\_2: 19.447386526571371  Rouge\_L: 23.93356175157701 |

**Bảng 4. 1: Kết quả Rouge-Score sau khi train model**

* **Điểm ROUGE-1**

**Điểm ROUGE-1 đo lường sự tương đồng về từ vựng giữa tóm tắt của mô hình và tóm tắt tham chiếu.**

**Precision của mô hình rất gần nhau giữa dữ liệu huấn luyện và kiểm tra, cho thấy mô hình có khả năng duy trì độ chính xác tốt trên cả hai tập dữ liệu.**

**Recall của mô hình cũng tương tự giữa hai tập dữ liệu, cho thấy mô hình có khả năng bao phủ các thông tin quan trọng tốt cả trên dữ liệu huấn luyện và kiểm tra,** nghĩa là mô hình không bỏ sót nhiều nội dung quan trọng.

**F-measure cũng gần nhau giữa hai tập dữ liệu, cho thấy sự cân bằng tốt giữa độ chính xác và độ hồi tưởng.**

* **Điểm ROUGE-2**

**Điểm ROUGE-2 đo lường sự tương đồng về các cụm từ hai từ giữa tóm tắt của mô hình và tóm tắt tham chiếu.**

**Precision, Recall, và F-measure cho ROUGE-2 đều tương đối thấp, cho thấy mô hình có thể chưa tốt lắm trong việc nắm bắt các cụm từ hai từ.**

**Sự khác biệt giữa huấn luyện và kiểm tra là rất nhỏ, cho thấy mô hình có khả năng tổng quát tương đối tốt trên các tập dữ liệu.**

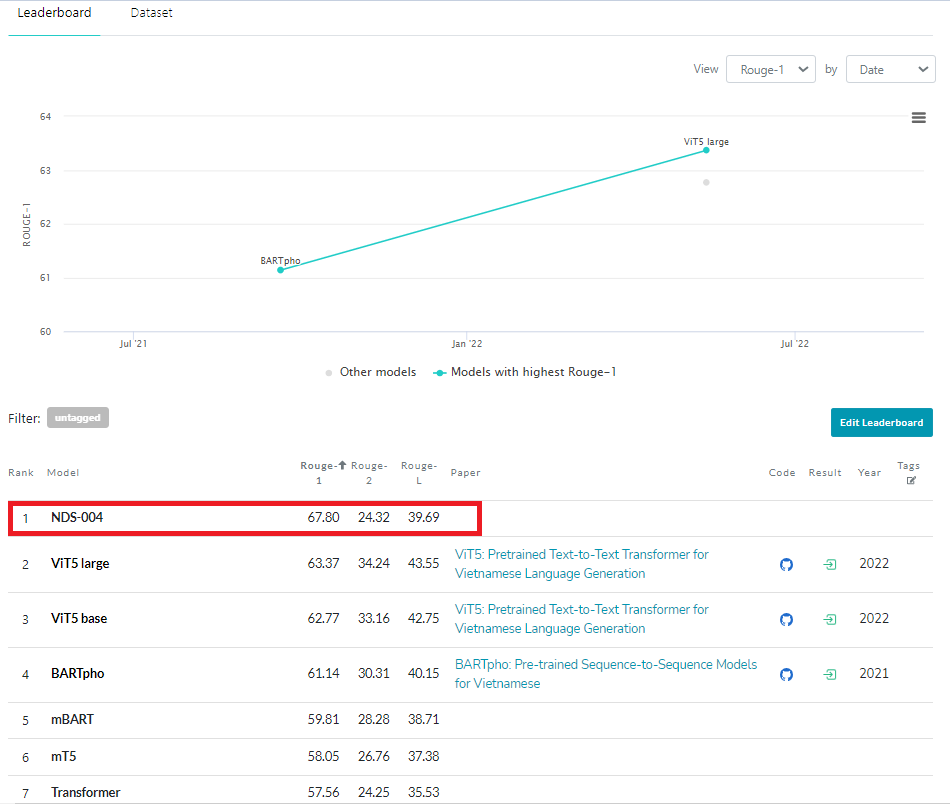
* **Điểm ROUGE-L**

**Điểm ROUGE-L đo lường sự tương đồng về chuỗi từ dài nhất giữa tóm tắt của mô hình và tóm tắt tham chiếu.**

**Precision, Recall, và F-measure cho ROUGE-L cho thấy mô hình có khả năng nắm bắt các chuỗi từ dài khá tốt.**

**Các điểm số rất gần nhau giữa huấn luyện và kiểm tra, cho thấy mô hình có khả năng tổng quát tốt trên các tập dữ liệu.**

So sánh với các phương pháp của các tác giả khác trên cùng tập dữ liệu:



* **Kết luận:**

Phương pháp được đề xuất có điểm **Rouge-1** và **Rouge-L** cao hơn so với các mô hình hiện đại như ViT5 large và BARTpho. Điều này cho thấy mô hình có khả năng tóm tắt văn bản tiếng Việt tốt.

* **Rouge-1** (67.8) cao hơn so với tất cả các mô hình khác, cho thấy mô hình có khả năng nắm bắt thông tin từ văn bản gốc rất tốt.
* **Rouge-2** (24.32) còn thấp, cho thấy mô hình chưa nắm bắt tốt mối quan hệ giữa các từ trong câu.
* **Rouge-L** (39.69) khá cao, cho thấy mô hình có khả năng giữ được cấu trúc dài hạn và độ mạch lạc của văn bản tốt.

**Phương pháp được đề xuất** đã cho thấy hiệu suất tốt, vượt trội hơn so với các mô hình hiện đại như ViT5 và BARTpho ở một số chỉ số. Điều này chứng tỏ rằng phương pháp được đề xuất có tiềm năng lớn trong việc tóm tắt văn bản tiếng Việt.

* **Hướng phát triển:**

Xem xét các kỹ thuật khác hoặc tinh chỉnh mô hình để cải thiện điểm ROUGE-2 và ROUGE-L, để nắm bắt tốt hơn các cụm từ và chuỗi từ dài hơn.

Cần cải thiện thêm các kỹ thuật xử lý ngôn ngữ và mô hình học máy để nâng cao điểm số ROUGE-2. Nghiên cứu và thử nghiệm các phương pháp biểu diễn văn bản và phân cụm khác để cải thiện chất lượng tóm tắt. Sử dụng thêm các nguồn dữ liệu khác nhau để đánh giá khả năng áp dụng của mô hình trên nhiều lĩnh vực khác nhau.

* **Một vài kết quả khi sử dụng mô hình để thực hiện tóm tắt văn bản:**

**Ví dụ 1:**

**Bản gốc:**

**theo báo long\_an , những ngày gần đây , nông\_dân trồng chanh ở huyện bến\_lức , tỉnh long\_an rất phấn\_khởi vì chanh được\_mùa , trúng giá . hiện\_tại , thương\_lái đến tận vườn mua chanh với giá 24.000 đồng / kg . nhiều thương\_lái cho biết , giá chanh tăng cao do đang vào mùa nắng\_nóng ,**

**nhu\_cầu sử\_dụng chanh làm nước giải\_khát tăng , đặc\_biệt là thị\_trường tp. hcm . cụ\_thể , chanh không hạt và có hạt được các thương\_lái thu\_mua tại vườn với giá dao\_động từ 23.000-25.000 đồng / kg , chanh bông tím có giá 24.000 đồng / kg . “ thị\_trường xuất\_khẩu chanh năm nay rộng hơn so năm\_ngoái nên giá chanh tăng và giữ được mức cao trong thời\_gian dài " , ông tám\_sơn ( chủ vựa chanh xã bình\_đức , huyện bến\_lức ) cho biết . theo nhiều người trồng chanh huyện bến\_lức , bình\_quân mỗi ha đất trồng chanh đạt năng\_suất hơn 20 tấn / năm . với mức giá hiện\_tại , sau khi trừ chi\_phí, người trồng chanh lãi vài trăm triệu đồng / ha / năm . nhiều nông\_dân cho biết , những năm trước , giá chanh thường tăng cao vài ngày lại giảm . tuy\_nhiên , năm nay , do thời\_tiết nắng\_nóng kéo\_dài nên giá chanh vẫn ở mức cao . còn theo khảo\_sát của pv báo vnexpress , tại các chợ trên địa\_bàn tp. hcm , giá chanh còn tăng cao hơn cả cam\_sành . chị lan , tiểu\_thương tại chợ xóm mới ( quận gò\_vấp , tp. hcm ) cho biết , giá chanh gần đây có lúc lên tới 45.000 đồng / kg với chanh không hạt , 35.000 đồng với chanh có hạt . như\_vậy , giá đã tăng gấp 4 lần so với cuối năm 2017 . chị kể , nếu bán\_lẻ thì cũng thu 3.000 đồng mỗi quả . “ mỗi ngày tôi lấy khoảng 5-10 kg nhưng đến chiều đã gần hết . giá tăng cao nhưng nắng\_nóng nên khách mua nhiều " , chị lan giải\_thích . chị vân , thương\_lái ở huyện giồng\_trôm ( tỉnh bến\_tre ) chia\_sẻ với tờ vnexpress , năm nay trồng chanh có lãi hơn cả cam . bởi lẽ , một ký cam\_sành loại ngon mua tại vườn chỉ 15.000 - 20.000 đồng / kg , còn chanh lên đến 30.000 đồng / kg . ngọc\_lài ( tổng\_hợp ) chanh tăng\_giá , nông\_dân tích\_cực chăm\_sóc ( ảnh : báo\_long\_an )**

**Bản tóm tắt:**

**theo báo long\_an những ngày gần đây nông\_dân trồng chanh ở huyện bến\_lức tỉnh long\_an rất phấn\_khởi vì chanh được\_mùa trúng giá chị lan tiểu\_thương tại chợ xóm mới quận gò\_vấp tp chị kể nếu bán\_lẻ thì cũng thu 3 000 đồng mỗi quả bởi lẽ một ký cam\_sành loại ngon mua tại vườn chỉ 15 000 20 000 đồng kg còn chanh lên đến 30 000 đồng kg'**

**Ví dụ 2:**

**Bản gốc:**

**sáng 2-2, trong chuyến làm việc tại tp.hcm, chủ tịch nước võ văn thưởng đã đến thăm, tặng quà, chúc tết nsnd kim cương và gs.bs nguyễn chấn hùng. đi cùng chủ tịch nước võ văn thưởng còn có ông lê khánh hải - chủ nhiệm văn phòng chủ tịch nước, ông nguyễn hồ hải - phó bí thư thường trực thành ủy tp.hcm. trao đổi với nsnd kim cương, chủ tịch nước võ văn thưởng tri ân những cống hiến, quá trình lao động của nghệ sĩ kim cương cho nền văn hóa của đất nước. không chỉ vậy, dù tuổi đã cao, bà vẫn luôn chăm lo, động viên đời sống vật chất, tinh thần cho các nghệ sĩ sân khấu, nhất là nghệ sĩ cao tuổi có hoàn cảnh khó khăn, không nơi nương tựa. nghệ sĩ kim cương còn có nhiều hoạt động hỗ trợ người khuyết tật và trẻ em mồ côi của tp.chủ tịch nước võ văn thưởng chúc nsnd kim cương năm mới nhiều niềm vui, may mắn, tiếp tục sống vui sống khỏe để truyền lửa cho thế hệ trẻ. thông qua nsnd kim cương, chủ tịch nước đã gửi lời thăm hỏi đến các văn nghệ sĩ khác tại tp. ông khẳng định song song với phát triển kinh tế, nước ta luôn chú trọng phát triển lĩnh vực văn hóa. trong đó, nhờ hoạt động nghệ thuật, qua các vai diễn, các văn nghệ sĩ đã góp phần rất lớn trong giáo dục truyền thống dân tộc, tuyên truyền chính sách của đảng và nhà nước. đáp lời, nsnd kim cương xúc động khi được lãnh đạo nhà nước quan tâm, thăm hỏi, động viên. bà kỳ vọng thế hệ trẻ với tinh thần nhiệt quyết, năng động sẽ có nhiều chính sách sáng tạo phát triển đất nước**

**Bản tóm tắt:**

**trong đó nhờ hoạt động nghệ thuật qua các vai diễn các văn nghệ sĩ đã góp phần rất lớn trong giáo dục truyền thống dân tộc tuyên truyền chính sách của đảng và nhà nước thông qua nsnd kim cương chủ tịch nước đã gửi lời thăm hỏi đến các văn nghệ sĩ khác tại tp bà kỳ vọng thế hệ trẻ với tinh thần nhiệt quyết năng động sẽ có nhiều chính sách sáng tạo phát triển đất nước sáng 2 2 trong chuyến làm việc tại tp hcm chủ tịch nước võ văn thưởng đã đến thăm tặng quà chúc tết nsnd kim cương và gs bs nguyễn chấn hùng**

# KẾT LUẬN

## Kết quả nghiên cứu của đề tài

Đề án này đề xuất một phương pháp tóm tắt văn bản tiếng Việt sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy.

**Mô hình đã chứng minh được khả năng tóm tắt văn bản tiếng Việt một cách hiệu quả và đồng nhất giữa dữ liệu huấn luyện và kiểm tra. Các điểm số ROUGE đạt được cho thấy mô hình có khả năng tổng quát tốt (khả năng duy trì hiệu suất trên dữ liệu kiểm tra).**

**Mô hình không chỉ hoạt động tốt trên dữ liệu huấn luyện mà còn duy trì hiệu suất tương tự trên dữ liệu kiểm tra, cho thấy mô hình có thể áp dụng cho các tập dữ liệu mới mà không bị giảm hiệu suất đáng kể.**

**Điểm ROUGE-1 cao hơn so với ROUGE-2 và ROUGE-L, điều này cho thấy mô hình có thể nắm bắt các từ đơn tốt hơn là các cụm từ và chuỗi từ dài hơn.**

## Hạn chế đề tài

**Chênh lệch giữa các chỉ số ROUGE:**

* Mặc dù kết quả Rouge-1 rất cao (67.8), nhưng Rouge-2 chỉ đạt 24.32. Điều này cho thấy mô hình có khả năng nhận diện từ và cụm từ đơn lẻ tốt, nhưng gặp khó khăn hơn trong việc nắm bắt mối quan hệ giữa các từ trong câu.

**Tính tổng quát của mô hình**:

* Hiệu suất mô hình trên tập dữ liệu huấn luyện và kiểm tra là tốt, nhưng có thể không đảm bảo rằng mô hình sẽ hoạt động hiệu quả trên các tập dữ liệu chưa từng thấy hoặc trong các ngữ cảnh thực tế.

**Thời gian build mô hình**:

* Thời gian xây dựng mô hình quá lâu trên tập dữ liệu lớn, dẫn đến việc thuê server mất nhiều chi phí. Điều này làm cho quá trình nghiên cứu và triển khai trở nên tốn kém và kém hiệu quả về mặt thời gian.

## Vấn đề kiến nghị và hướng đi tiếp theo của đề tài

Đối với các nghiên cứu trong tương lai, đề án có thể cải tiến bằng cách xem xét các vấn đề sau:

* Tiếp tục tối ưu hóa mô hình hiện tại và thử nghiệm trên các bộ dữ liệu khác nhau để đánh giá khả năng tổng quát của mô hình.
* Sử dụng các kỹ thuật tiên tiến hơn như attention mechanisms hoặc transformer architectures để nắm bắt tốt hơn mối quan hệ giữa các từ trong câu.
* Xem xét các kỹ thuật mới trong lĩnh vực xử lý ngôn ngữ tự nhiên để tiếp tục nâng cao chất lượng tóm tắt.
* Nghiên cứu và áp dụng các kỹ thuật giảm thiểu độ phức tạp của mô hình như pruning, quantization hoặc sử dụng các mô hình nhỏ gọn hơn mà vẫn duy trì hiệu suất tốt.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Hoang Anh Pham. (2018). Vietnamese doc summarization basic. Available from:

<https://github.com/hoanganhpham1006/Vietnamese_doc_summarization_basic>

1. Xin Rong, 2016. word2vec Parameter Learning Explained. Available from: https://arxiv.org/abs/1411.2738
2. Facebook AI Research. fastText Vietnamese word vectors full. Available from: <https://www.kaggle.com/datasets/aeryss/fasttext-vietnamese-word-vectors-full>
3. Phạm Hữu Quang. (2018). Xây dựng mô hình không gian vector cho Tiếng Việt. Available from:

<https://github.com/hoanganhpham1006/Vietnamese_doc_summarization_basic>

1. Nguyen, Van-Hau & Nguyen, Thanh-Chinh & Nguyen, Minh-Tien & Hoai, Nguyen. (2019). VNDS: A Vietnamese Dataset for Summarization. 375-380. 10.1109/NICS48868.2019.9023886.
2. Nguyen, Dat Quoc & Nguyen, Anh. (2020). PhoBERT: Pre-trained language models for Vietnamese. 1037-1042. 10.18653/v1/2020.findings-emnlp.92.
3. H. Q. To, and K. V. Nguyen, and N. L. -T. Nguyen, and A. G. T. Nguyen, Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization, in Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, Shanghai: China, pp. 692-699, 2021.
4. Hieu Nguyen, Long Phan, James Anibal, Alec Peltekian, and Hieu Tran. 2021. Viesum: How robust are transformer-based models on vietnamese summarization?
5. T. A. Nguyen-Hoang and K. Nguyen and Q. V. Tran, TSGVi: A graphbased summarization system for Vietnamese documents, Journal of Ambient Intelligence and Humanized Computing, vol. 3, no. 4, pp. 305-313, Jun., 2012. DOI: 10.1007/s12652-012-0143-x.
6. D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” in Findings of EMNLP, 2020.
7. [13]. D. Q. Thang, L. H. Phuong, N. T. M. Huyen, N. C. Tu, M. Rossignol, and V. X. Luong, “Word segmentation of Vietnamese texts: a comparison of approaches,” in LREC, 2008.
8. Minh-Tien Nguyen, Hoang-Diep Nguyen, Thi-HaiNang Nguyen, and Van-Hau Nguyen. 2018. Towards state-of-the-art baselines for vietnamese multi-document summarization. In 2018 10th International Conference on Knowledge and Systems Engineering (KSE), pages 85–90.
9. Rothe, Sascha & Narayan, Shashi & Severyn, Aliaksei. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. Transactions of the Association for Computational Linguistics. 8. 264-280. 10.1162/tacl\_a\_00313.
10. M. Allahyari, and S. Pouriyeh, and M. Assefi, and S. Safaei, and E. D. Trippe, and J. B. Gutierrez, and K. Kochut, Text summarization techniques: A brief survey, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 10, pp. 397-405, 2017. DOI: 10.14569/IJACSA.2017.081052.
11. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer,” in NAACL, 2021.
12. T. V. Bui, T. O. Tran, and P. Le-Hong, “Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models,” in PACLIC, 2020.
13. T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, “VnCoreNLP: A Vietnamese Natural Language Processing Toolkit,” in NAACL (Demonstrations), 2018.
14. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B.Gutierrez, and K. Kochut, “*Text summarization techniques: A briefsurvey,” International Journal of Advanced Computer Science andApplications (IJACSA),* vol. 8, no. 10, pp. 397-405, 2017. DOI:10.14569/IJACSA.2017.081052. (PDF) Text Summarization on Large-scale Vietnamese Datasets. Available from: https://www.researchgate.net/publication/366806176\_Text\_Summarization\_on\_Large-scale\_Vietnamese\_Datasets [accessed Nov 03 2023].
15. A. Nguyen-Hoang, K. Nguyen, and Q. V. Tran, “TSGVi: A graph-based summarization system for Vietnamese documents,” *Journal ofAmbient Intelligence and Humanized Computing*, vol. 3, no. 4, pp.305-313, Jun. 2012. DOI: 10.1007/s12652-012-0143-x. (PDF) Text Summarization on Large-scale Vietnamese Datasets. Available from: https://www.researchgate.net/publication/366806176\_Text\_Summarization\_on\_Large-scale\_Vietnamese\_Datasets [accessed Nov 03 2023].
16. Dr Inderjeet Mani , Mark T Maybury. (2018). Advances in Automatic Text Summarization (Mit Press).
17. Hong Phuong L., Thi Minh Huyen N., Roussanaly A., Vinh H.T. (2008) A Hybrid Approach to Word Segmentation of Vietnamese Texts. In: Martín-Vide C., Otto F., Fernau H. (eds) Language and Automata Theory and Applications. LATA 2008. Lecture Notes in Computer Science, vol 5196. Springer, Berlin, Heidelberg.
18. Nguyễn Thị Ngọc Tú, Nguyễn Thị Thu Hà, Lê Thanh Hương, Hồ Ngọc Vinh, Đào Thanh Tĩnh, Nguyễn Ngọc Cương. “Ứng dụng mô hình đồ thị trong tóm tắt đa văn bản tiếng Việt”. Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015.
19. Sam Wiseman and Alexander M. Rush, 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. School of Engineering and Applied Sciences Harvard University Cambridge, MA, USA-In arXiv:1606.02960v2 [cs.CL].
20. Sam Wiseman and Alexander M. Rush, 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. School of Engineering and Applied Sciences Harvard University Cambridge, MA, USA-In arXiv:1606.02960v2 [cs.CL].
21. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013. Efficient Estimation of Word Representations in Vector Space – In arXiv:1301.3781v3 [cs.CL].
22. Trương Quốc Định, Nguyễn Quang Dũng. “Một giải pháp tóm tắt văn bản tiếng Việt tự động”. Hội thảo quốc gia lần thứ XV: một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông Hà Nội 03-04/12/2012.