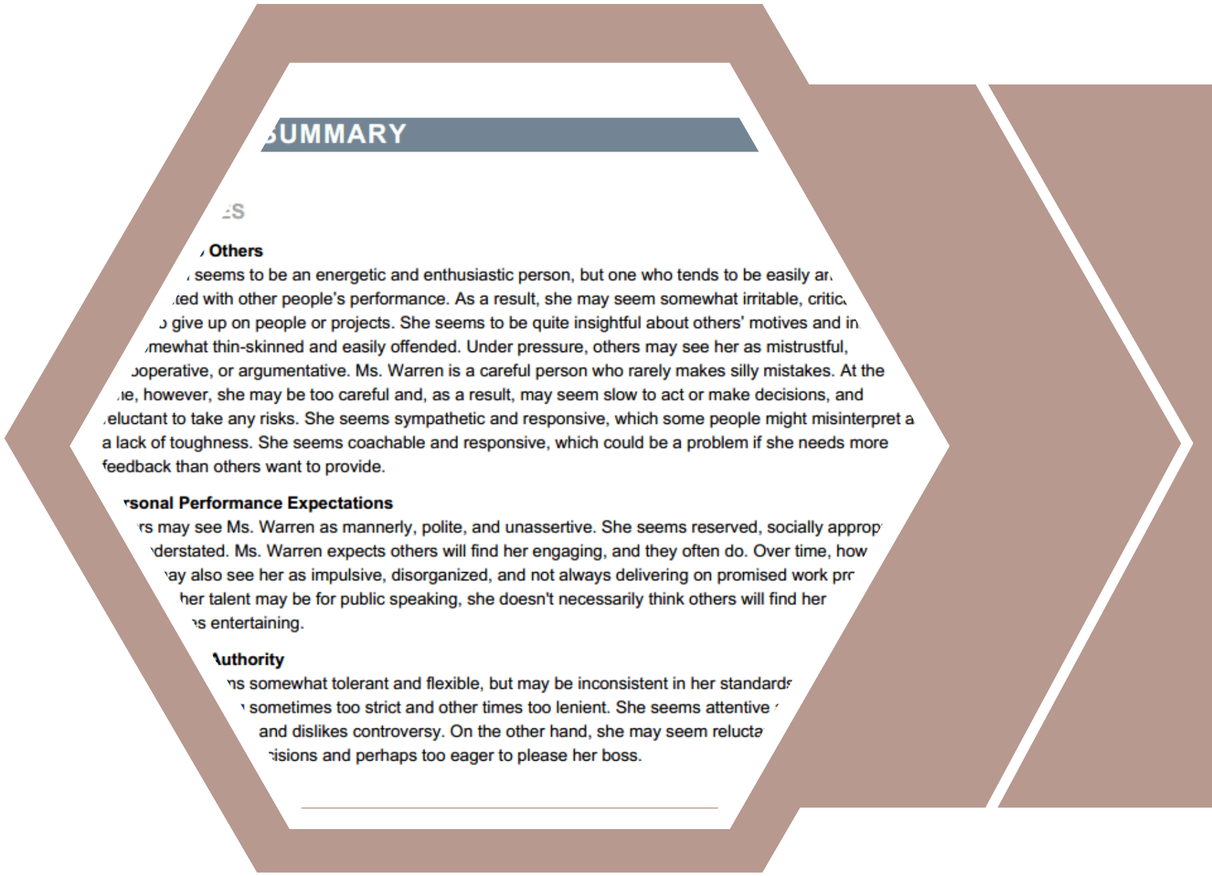
XÂY DỰNG MÔ HÌNH TÓM TẮT VĂN BẢN TIẾNG VIỆT

Báo Cáo Cuối Kỳ



GIẢNG VIÊN : TS. PHẠM TIẾN LÂM

THS. ĐẶNG VĂN BÁU

THÀNH VIÊN NHÓM 10

Đỗ Minh Tuấn - 20002175

Lã Anh Trúc - 20002169

Lê Hồng Thạch - 20002162

Trần Kim Phượng - 20002154

Phạm Hoàng An - 20002102

TỔNG QUAN

- I. Giới thiệu bài toán
- II. Giải quyết bài toán
- III. Mô phỏng và thực hiện
- IV. Kết luận và đánh giá

Lý do chọn đề tài

- Đề tài "Tóm tắt văn bản tiếng Việt" giải quyết nhu cầu ngày càng tăng về tóm tắt văn bản tiếng Việt, đặc biệt trong lĩnh vực tin tức, nghiên cứu khoa học và kinh doanh.
- Sử dụng học máy để tóm tắt giúp tiết kiệm thời gian và dễ dàng tiếp cận thông tin.
- Tóm tắt văn bản tiếng Việt là lĩnh vực thú vị và có tính ứng dụng cao, đang được quan tâm và nghiên cứu rộng rãi.

I. Giới thiệu bài toán

Với đầu vào là một văn bản gồm N câu, mô hình sẽ loại bỏ đi những câu có ý nghĩa tương tự nhau hay ko phải ý chính để tạo ra một văn bản tóm tắt gồm K câu và giữ lại được ý chính của đoạn.

II. Giải quyết bài toán

1 Cách tiếp cận

2 Một số phương pháp sử dụng

3 Thực hiện

2.1. Tiếp cận bài toán

Có hai phương pháp chính thường được sử dụng để tóm tắt văn bản trong Xử Lý Ngôn Ngữ Tự Nhiên (NLP):

1. Extraction-based (phương pháp dựa trên trích xuất)
các thông tin quan trọng, các cụm từ chính (key phrase)được trích xuất ra từ văn bản ban đầu l từ tài liệu và kết hợp chúng để tạo ra một bản tóm tắt

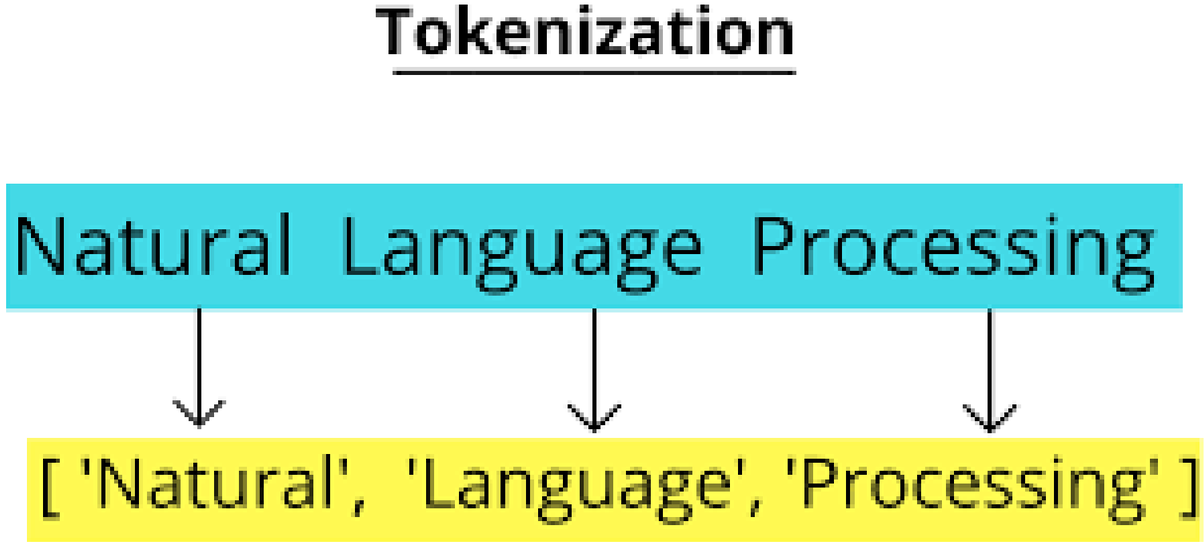
2. Abstraction-based (phương pháp dựa trên trừu tượng)
Chọn các từ dựa trên sự hiểu biết ngữ nghĩa, ngay cả những từ đó không xuất hiện trong các tài liệu gốc.

2.2. Các phương pháp và kỹ thuật

+ Tokenization
+ Word2Vec
+ Long Short Term Memory (LSTM)
+ Autoencoder

Tonkenization

- Tokenization (tách từ) là một trong những bước quan trọng nhất trong quá trình tiền xử lý văn bản, đó là quá trình tách một cụm từ, câu, đoạn văn, một hoặc nhiều tài liệu văn bản thành các đơn vị nhỏ hơn.
- Mỗi đơn vị nhỏ hơn này được gọi là Tokens.



Phân loại

Mã hoá dựa trên từ
(word-based tokenization algorithm)

Mã hoá dựa trên từ phụ
(subword-based tokenization algorithm)

Mã hoá dựa trên kí tự
(character-based tokenization algorithm)

Word2vec

Word2vec là một mô hình đơn giản giúp tạo ra các biểu diễn embedding của từ trong một không gian có số chiều thấp hơn nhiều lần so với số từ trong từ điển. Word2Vec được xây dựng bằng cách phân tích ngữ cảnh xuất hiện của từ trong văn bản.

Ý tưởng cơ bản của word2vec:

- Hai từ xuất hiện trong những văn cảnh giống nhau thường có ý nghĩa gần với nhau.
- Ta có thể đoán được một từ nếu biết các từ xung quanh nó trong câu.

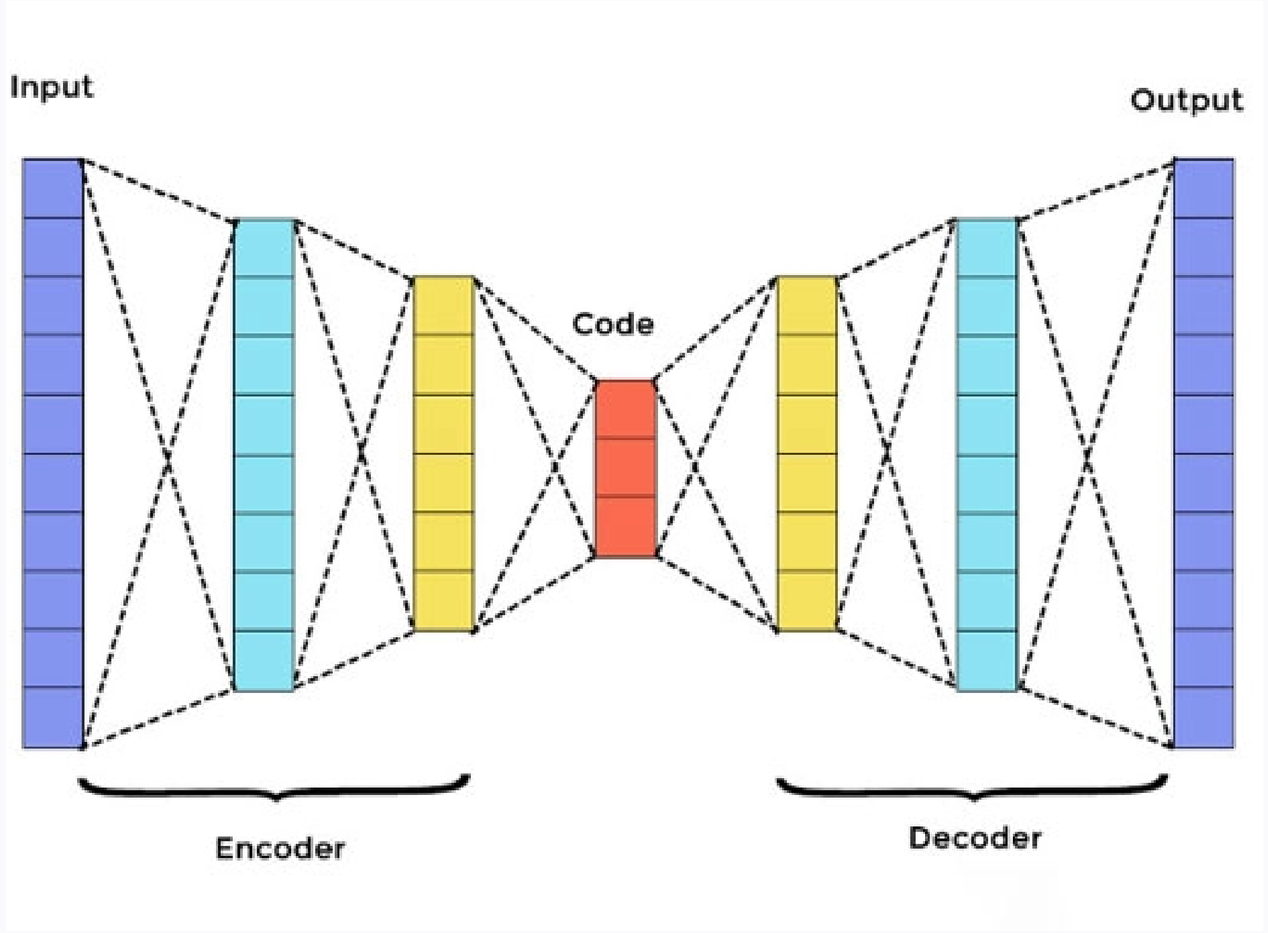
Autoencoder

- Mô hình autoencoder là một trong những mô hình học sâu (deep learning) phổ biến trong lĩnh vực xử lý ảnh và xử lý ngôn ngữ tự nhiên. Autoencoder là một loại mạng neural nhân tạo được sử dụng để học các loại mã hóa dữ liệu không giám sát (unsupervised learning).
- Mục đích của Autoencoder là học cách biểu diễn chiều nhỏ hơn (mã hóa) cho dữ liệu có chiều cao hơn. Đây cũng là lý do mà Autoencoder thường được dùng cho các bài toán giảm chiều dữ liệu hay trích xuất đặc trưng.

Kiến trúc của Autoencoder

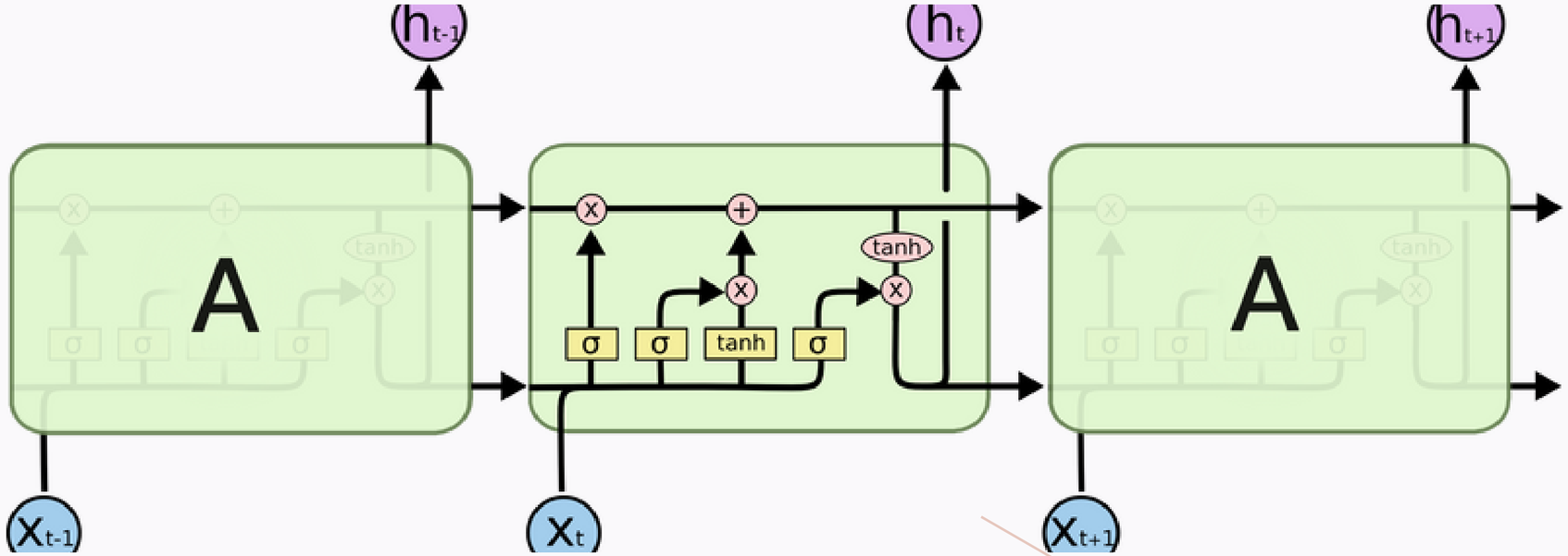
Kiến trúc của Autoencoder bao gồm 3 phần chính:

- Encoder: Module có nhiệm vụ nén dữ liệu đầu vào thành một biểu diễn được mã hóa (coding) .
- Bottleneck: Module chứa các biểu diễn tri thức được nén (chính là output của Encoder) .
- Decoder: Module giúp mạng giải nén các biểu diễn tri thức và tái cấu trúc lại dữ liệu từ dạng mã hóa của nó, mô hình học dựa trên việc so sánh đầu ra với đầu vào ban đầu của Encoder.

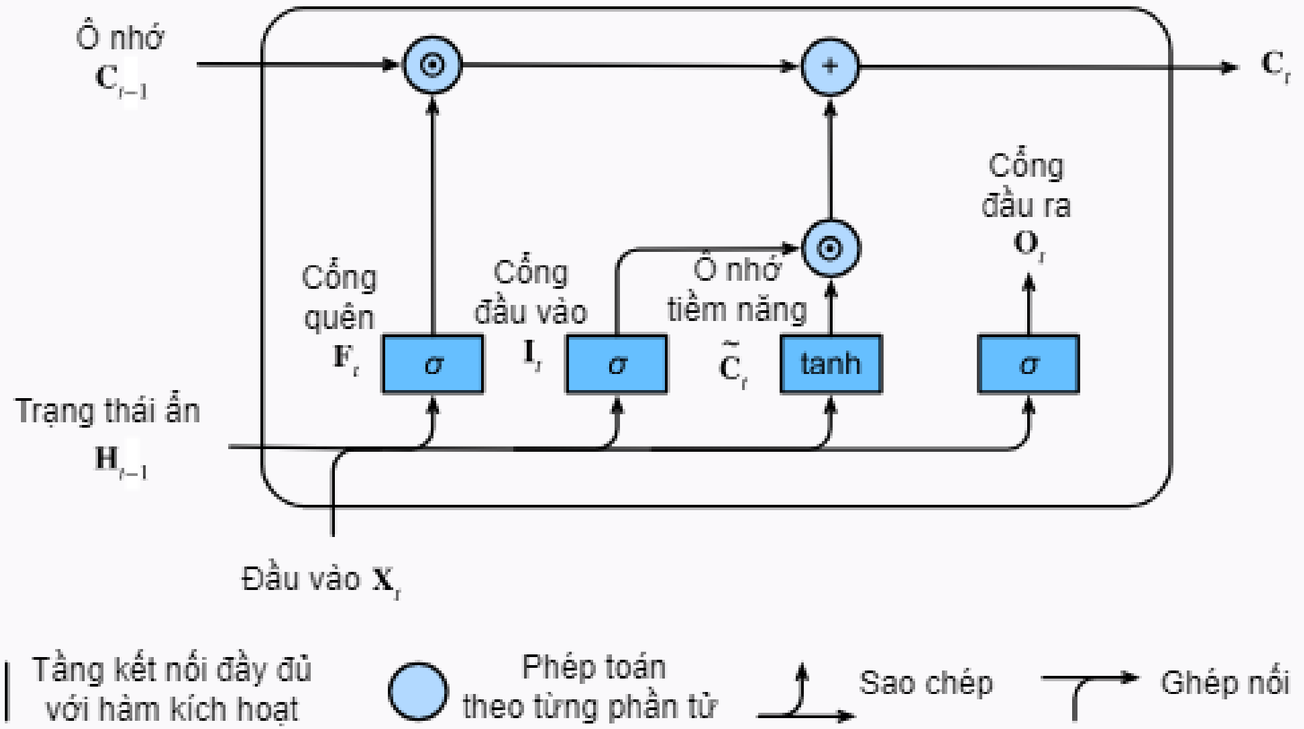


LSTM

- LSTM là một kiến trúc mạng nơ-ron đặc biệt được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên (NLP) và các bài toán chuỗi thời gian (time series).
- LSTM được thiết kế để giải quyết vấn đề biến mất gradient trong quá trình huấn luyện mạng nơ-ron sâu.



- LSTM có khả năng lưu trữ thông tin lâu dài và xử lý các chuỗi dữ liệu dài và phức tạp hơn so với các kiến trúc mạng nơ-ron tái phát khác
- Cơ chế hoạt động của LSTM là chỉ ghi nhớ những thông tin liên quan, quan trọng cho việc dự đoán, còn các thông tin khác sẽ được bỏ đi.



Các bước thực hiện

Theo phương pháp trích dẫn

Chuẩn bị dữ liệu → Xử lý dữ liệu → Tokenized thành các câu → Tìm tần suất trọng số → Tìm ngưỡng giá trị → Nhận bản tóm tắt → **Thu được văn bản tóm tắt**

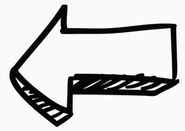
Theo phương pháp trừu tượng

Chuẩn bị dữ liệu

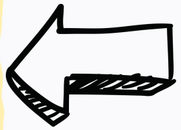
Tiền xử lý văn bản

Encoder

LSTM layer

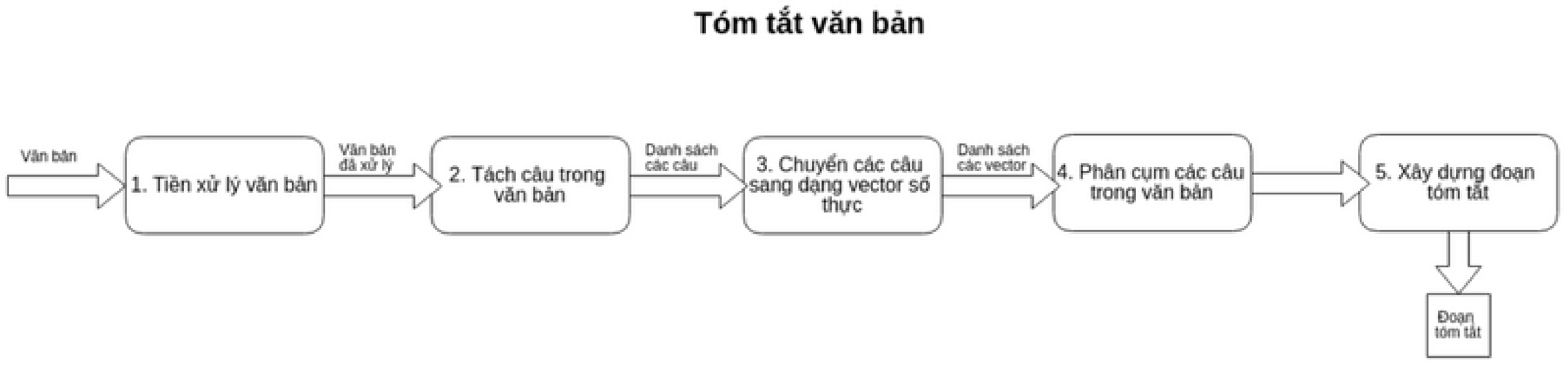


Decoder



Bản tóm tắt

Tóm tắt văn bản bằng phương pháp phân cụm



III. Mô phỏng thực hiện bài toán

1.850K

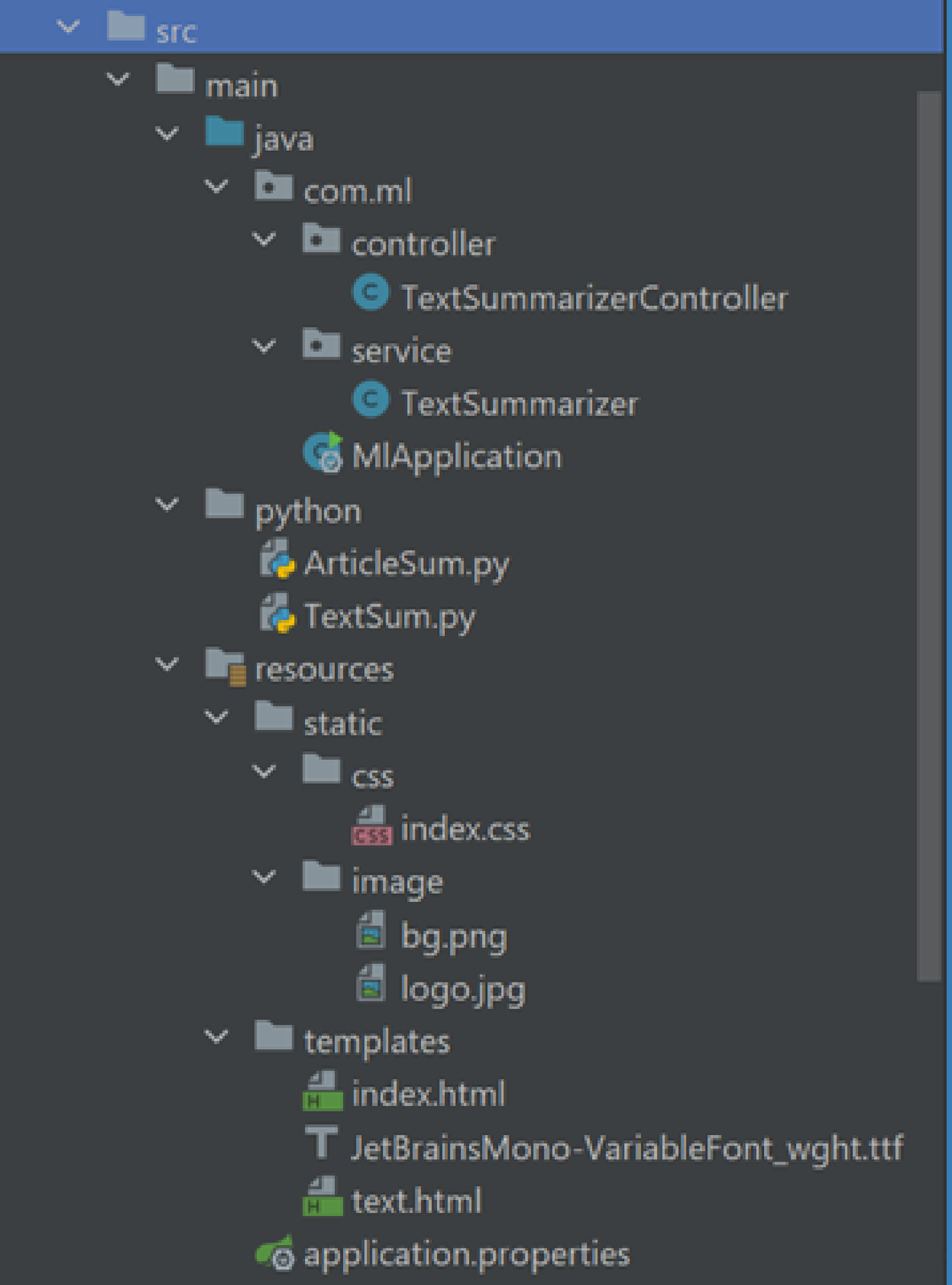
1 Giới thiệu tổng quan về trang web

2 Chức năng

3 Demo

Tổng quan - Phương pháp làm việc

Công nghệ sử dụng:



Cây thư mục Project

2 chức năng chính

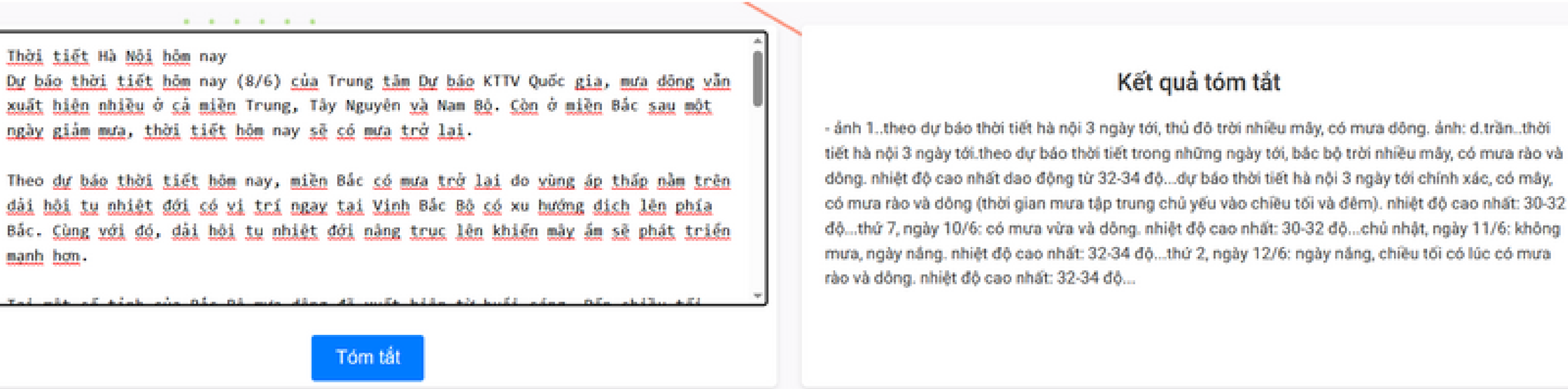
1. **Tóm tắt bài báo Tiếng Việt qua URL.**
2. **Tóm tắt văn bản Tiếng Việt tùy chọn.**

1.850K

DEMO

IV. Kết quả đạt được và đánh giá

Hình 4.1: Tóm tắt bài báo Tiếng Việt qua URL



Hình 4.2: Tóm tắt văn bản Tiếng Việt

Đánh giá

Ưu điểm

- Tốc độ tóm tắt nhanh hơn nhiều so với cách tóm tắt thủ công.
- Độ chính xác khá ổn, cho ra được kết quả là đoạn văn tóm tắt.
- Có giao diện web thân thiện với người dùng,dễ dàng sử dụng.
- Người dùng có thể tùy chọn nhu cầu tóm tắt cả bài hay một phần văn bản.

Nhược điểm

- Độ chính xác: Chưa hoàn toàn chính xác nên có thể dẫn đến sự sai sót và hiểu nhầm trong quá trình tóm tắt.
- Dữ liệu huấn luyện vẫn chưa đủ lớn, chưa đại diện cho tất cả loại văn bản
- Khó khăn trong việc hiểu các cấu trúc ngôn ngữ phức tạp do còn thiếu những kiến thức xử lý ngôn ngữ tự nhiên.
- Các tài liệu về xử lí ngôn ngữ tiếng Việt còn hạn chế

Kết Luận

- Nhóm đã tìm hiểu về ứng dụng của học máy kết hợp với xử lý ngôn ngữ tự nhiên (NLP) đê tóm tắt văn bản tiếng việt qua hai phương pháp chính và các kỹ thuật .
- Nghiên cứu và thực nghiệm nhiều phương pháp khác nhau để thực hiện trong đó có phương pháp trích dẫn, phương pháp trừu tượng và phương pháp phân cụm.
- Tuy nhiên, chúng tôi vẫn còn gặp những khó khăn và hạn chế khi thực hiện bài tập lớn này. Xử lý ngôn ngữ tự nhiên (NLP) cũng là một trong những lĩnh vực còn nhiều khó khăn với chúng tôi để tiếp cận và xử lý đạt kết quả tốt nhất.

Tài liệu tham khảo

[1]. Comprehensive Guide to Text Summarization using Deep Learning in Python(analyticsvidhya.com)

[2].https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning

[3].https://viblo.asia/p/xay-dung-chuong-trinh-tom-tat-van-ban-tieng-viet-don-gian-voi-machine-learning

[4].https://hoctructuyen123.net/tom-tat-van-ban-trong-hoc-may/?fbclid=IwAR2g61kb-19-8hbxkrCwt9KEI7hDBsYQi2mNX3orbsVGfuzDqMl5h4a_f50

[5].Word2vec:https://machinelearningcoban.com/tabml_book/ch_embedding/word2vec.html

[6]. LSTM:https://websitehcm.com/long-short-term-memory-lstm-la-gi/

[7].Autoencoder:https://bizflycloud.vn/tin-tuc/autoencoder-la-gi-20220526165157229.htm

THANK YOU