# TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

# TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

# KHOA CÔNG NGHỆ THÔNG TIN



# ĐỒ ÁN CUỐI KÌ

# NHẬP MÔN HỌC MÁY

# Tóm tắt văn bản tiếng việt tự động

*Người hướng dẫn*: **TS BÙI THÀNH HÙNG**

*Người thực hiện*: **NGUYỄN MINH NHỰT – 518H0545**

**ĐẶNG THỌ DUY – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

**Tóm tắt văn bản tiếng việt tự động**

*Người hướng dẫn*: **TS BÙI THÀNH HÙNG**

*Người thực hiện*: **NGUYỄN MINH NHỰT – 518H0545**

**ĐẶNG THỌ DUY – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

# LỜI CẢM ƠN

Lời đầu tiên, chúng chúng em xin được gửi lời cảm ơn chân thành đến với Khoa CNTT. Chúng em xin chân thành cảm ơn sự hỗ trợ, giúp đỡ trong quá trình thực hiện bài báo cáo môn Nhập môn học máy này.

Chúng em xin chân thành cảm ơn thầy Bùi Thanh Hùng - thầy đã trực tiếp hướng dẫn và chỉ bảo chúng chúng em hoàn thành bài báo cáo này.

Chúng tôi xin chân thành cảm ơn những người bạn đang học tập và làm việc tại trường Đại học Tôn Đức Thắng và gia đình đã đồng hành, động viên, tạo điều kiện và giúp đỡ chúng em trong suốt quá trình này.

Do năng lực thực tế của chúng em còn non yếu nên chúng em đảm bảo còn nhiều thiếu sót, mong quý thầy cô bỏ qua. Đồng thời chúng em mong nhận được nhiều ý kiến đóng góp từ nhiều nguồn khác nữa, điều này sẽ giúp chúng em tích lũy thêm kinh nghiệm để hoàn thành bài báo cáo tốt nghiệp sắp tới đạt kết quả tốt hơn.

# ĐỒ ÁN ĐƯỢC HOÀN THÀNH

# TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của TS Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 21 tháng 04 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Minh Nhựt*

*Đặng Thọ Duy*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

Phần I: Phân tích yêu cầu bài toán: Phân tích được yêu cầu của bài toán là gì (1 điểm)

Phần II: Phương pháp giải quyết: Trình bày được các phương pháp giải quyết bài toán, Giải

thích lý do tại sao chọn phương pháp này, Vẽ được sơ đồ tổng quát giải quyết bài

toán (2 điểm)

Phần III: Hiện thực - Viết code theo phương pháp giải quyết ở trên: Trình bày được cụ thể giải

thuật sử dụng để giải quyết bài toán (Lưu đồ giải thuật), các tham số sử dụng, các thư

viện sử dụng, code của bài toán … (4.5 điểm)

Phần IV: Đánh giá kết quả đạt được: So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ

so sánh giữa các phương pháp theo các độ đo ví dụ như: Accuracy, MSE, RMSE,

MAP, …. (hãy lựa chọn ít nhất 2 độ đo trong các độ đo phổ biến để đánh giá bài toán

trên) (1.5 điểm)

Hướng phát triển trong tương lai: Đưa ra được hướng phát triển trong tương lai và

giải thích lý do tại sao lại đưa ra hướng phát triển đó (0.25 điểm).

# MỤC LỤC

[TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM 1](#_Toc70152196)

[TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG 1](#_Toc70152197)

[KHOA CÔNG NGHỆ THÔNG TIN 1](#_Toc70152198)

[ĐỒ ÁN CUỐI KÌ 1](#_Toc70152199)

[NHẬP MÔN HỌC MÁY 1](#_Toc70152200)

[Tóm tắt văn bản tiếng việt tự động 1](#_Toc70152201)

[LỜI CẢM ƠN i](#_Toc70152202)

[ĐỒ ÁN ĐƯỢC HOÀN THÀNH ii](#_Toc70152203)

[TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii](#_Toc70152204)

[TÓM TẮT iv](#_Toc70152205)

[MỤC LỤC 1](#_Toc70152206)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 2](#_Toc70152207)

[DANH MỤC CÁC HÌNH VẼ 3](#_Toc70152208)

[DANH MỤC CÁC BẢNG 4](#_Toc70152209)

[PHẦN I 5](#_Toc70152210)

[PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN 5](#_Toc70152211)

[1..1 Giới thiệu về bài toán 5](#_Toc70152212)

[1.2 Phân tích yêu cầu của bài toán 5](#_Toc70152213)

[1.2.1 Khái quát về bài toán 5](#_Toc70152214)

[1.2.2 Yêu cầu của bài toán 6](#_Toc70152215)

[PHẦN II 8](#_Toc70152216)

[PHƯƠNG PHÁP GIẢI QUYẾT 8](#_Toc70152217)

[2.1 Các phương pháp giải quyết bài toán 8](#_Toc70152218)

[2.1.1 Hướng trích chọn – Extraction 8](#_Toc70152219)

[2.1.2 Hướng tóm lược – Abstraction 9](#_Toc70152220)

[2.2 Phương pháp đề xuất giải quyết bài toán 10](#_Toc70152221)

[2.3 Phương pháp giải quyết bài toán 11](#_Toc70152222)

[2.3.1 Mô hình tổng quát 11](#_Toc70152223)

[2.3.2 Đặc trưng của mô hình đề xuất 12](#_Toc70152224)

[PHẦN III 19](#_Toc70152225)

[THỰC NGHIỆM 19](#_Toc70152226)

[3.1 Dữ liệu 19](#_Toc70152227)

[3.1.1 Nguồn gốc dữ liệu 19](#_Toc70152228)

[3.1.2 Mô tả dữ liệu 19](#_Toc70152229)

[3.2 Xử lý dữ liệu 21](#_Toc70152230)

[3.3 Công nghệ sử dụng 21](#_Toc70152231)

[3.4 Cách đánh giá 22](#_Toc70152232)

[3.5 Kết quả đạt được 22](#_Toc70152233)

[PHẦN IV 24](#_Toc70152234)

[KẾT LUẬN 24](#_Toc70152235)

[TÀI LIỆU THAM KHẢO 26](#_Toc70152236)

[PHỤ LỤC 27](#_Toc70152237)

[TỰ ĐÁNH GIÁ 28](#_Toc70152238)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

CSTD Công suất tác dụng

MF Máy phát điện

BER Tỷ lệ bít lỗi

# DANH MỤC CÁC HÌNH VẼ

[Hình 2.1: Kiến trúc FTP 1](#_Toc387689394)

Hình 2.2: Kiến trúc TP 1

# DANH MỤC CÁC BẢNG

[Bảng 3.1 Ví dụ cho chèn bảng 1](#_Toc387689363)

[Bảng 3.2 Kết quả 1](#_Toc387689363)

# PHẦN I

# PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN

## 1..1 Giới thiệu về bài toán

Ngày nay công nghệ thông tin và máy tính phát triển mạnh mẽ đi theo đó là khối lượng tri thức, 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 các thông tin quan trọng trong khối tri thức đồ sộ này tốn rất nhiều thời gian và chi phí (đặc biệt là chi phí cho hạ tầng và truyền dẫn thông tin đáp ứng yêu cầu cho một số lượng ngày càng nhiều các thiết bị cầm tay). Để tăng hiệu quả trong việc tiếp nhận thông tin ấy, nhiều nghiên cứu về khai phá dữ liệu và xử lý ngôn ngữ tự nhiên đã được thực hiện. Một trong những nghiên cứu quan trọng đóng vai trò then chốt đó tóm tắt văn bản tự động.

Hiện trên thị trường đã có nhiều mô hình thực hiện và giải quyết bài toán tóm tắt văn bản nhưng đa số đó là văn bản tiếng Anh vì thế tóm tắt văn bản nói riêng và xử lý ngôn ngữ tự nhiên nói chung áp dụng cho tiếng Việt gặp nhiều thách thức hơn. Sở dĩ là vì tiếng Việt với đặc trưng như là tiếng đơn âm và có thanh điệu nên việc tách từ, tách các

thành phần ngữ nghĩa trong câu tiếng Việt đòi hỏi xử lý phức tạp hơn so với xử lý câu trong tiếng Anh, song số lượng kho dữ liệu (dataset) bằng tiếng Việt được chuẩn hóa còn hạn hẹp và tính chính thức cũng chưa cao như tiếng Anh.

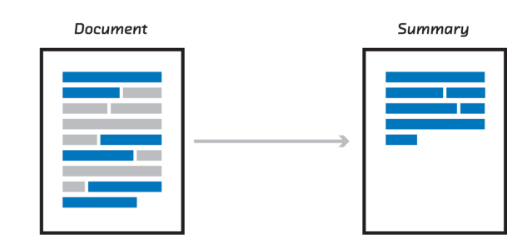
## 1.2 Phân tích yêu cầu của bài toán

### 1.2.1 Khái quát về bài toán

Đầu tiên, tóm tắt được định nghĩa như là một văn bản dược 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.

### 1.2.2 Yêu cầu của bài toán

Tất cả những gì chúng ta cần cho bài toán này là 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.



#### Hình 1.1: Yêu cầu của bài toán tóm tắt văn bản

Mục tiêu chính của bài toán là tạo ra bản tóm tắt giống như cách con người tóm tắt, vì thế mà đây là bài toán được đánh giá là đầy thách thức. Khi chính con ngườ thực hiện tóm tắt một văn bản, chúng ta càn phải đọc toàn bộ nội dung rồi dựa trên sự hiểu biết và tiếp thu của bản thân để viết lại một đoạn tóm tắt nhằm làm nổi bật các ý chính của văn bản gốc.

Và chúng ta cũng đã biết, tất cả những gì máy tính hiểu là dãy bit 0 và 1 vì thế nó khó có thể có được tri thức và khả năng cảm thụ ngôn ngữ như của con người, nên việc thực hiện tóm tắt văn bản tự động là một công việc phức tạp.

# PHẦN II

# PHƯƠNG PHÁP GIẢI QUYẾT

## 2.1 Các phương pháp giải quyết bài toán

Theo khảo sát của chúng em, ở đây có khá nhiều hướng để tiếp cận cho bài toán tóm tắt văn bản tiếng việt. Theo tác giả Nguyễn Mạnh Cường [3], thầy đã đề ra các giải pháp như sau:

* Tóm tắt theo hướng trích chọn – extractive based summarization
* Tóm tắt theo hướng tóm lược – abstractive based summarization
* Tóm tắt bằng đồ thị hóa
* Tóm tắt bằng các mô hình học máy

Nhưng chung quy, chúng ta có 2 hướng chính để ta giải quyết bài toán này, đầu tiên là giải quyết theo hướng trích chọn – extraction và tóm lược – abstraction.

### 2.1.1 Hướng trích chọn – Extraction

Đầu tiên là giải quyết theo hướng trích chọn – extraction, chúng ta sẽ thực hiện đánh giá các phần quan trọng của văn bản, 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 hay được gọi là khối lượng từ trọng tâm – Weight text.

Nói một cách dễ hiểu hơn là hướng tóm lược đóng vai trò như 1 cây viết dạ quang – highlighter, thứ mà sẽ làm nổi bật lên những ý chính trong văn bản gốc



#### Hình 2.1: Highlighter = Hướng trích chọn

Ví dụ như sau:

**Văn bản gốc**: Peter và Elizabeth bắt taxi để đi dự buổi tiệc tối nay ở thành phố Duo Flora. Trong khi đang dự tiệc, Elizabeth đột nhiên bất tỉnh và phải nhập viện khẩn cấp.

**Văn bản tóm tắt**: Peter và Elizabeth dự buổi tiệc ở thành phố Duo Flora, Elizabeth nhập viện khẩn cấp

#### **2.1.1.1 Phương pháp minh họa:**

Như trích dẫn bên dưới, mục [8], tạm dịch là “Tóm tắt hướng trích chọn dựa trên trích xuất sự kiện” của nhóm tác giả Maofu Liu, Wenje Li, Mingli Wu, Quin Lu.

Khái quát thì tóm tắt sự kiện là trích xuất và tố chức các đoạn tóm tắt theo các điều kiện của sự kiện có trong câu muốn tóm tắt. Cụ thể là chúng ta sẽ

### 2.1.2 Hướng tóm lược – Abstraction

Phương pháp này nhằm tạo ra các văn bản tóm lược có thể không bao gồm các từ hay cụm từ trong văn bản gốc. Khi thực hiện phương pháp này, chúng ta phải làm máy cố gắng hiểu và đánh giá văn bản bằng cách sử dụng những các kỹ thuật xử lý ngôn ngữ tự nhiên để tạo ra 1 văn bản tóm tắt càng ngắn càng tốt nhưng vẫn có thể truyền đạt được những thông tin quan trọng nhất từ văn bản gốc.

Nói một cách dễ hiểu hơn là hướng tóm lược đóng vai trò như 1 cây viết bình thường, thứ mà sẽ viết ra những thứ dựa trên văn bản gốc nhưng chưa có sẵn trong đó.



#### Hình 2.2: Viết thường = Hướng tóm lược

Ví dụ như sau:

**Văn bản gốc**: **Peter** và **Elizabeth** bắt taxi để đi **dự** **buổi** **tiệc** tối nay ở thành phố Duo Flora. Trong khi đang dự tiệc, **Elizabeth** đột nhiên bất tỉnh và phải **nhập** **viện** khẩn cấp.

**Văn bản tóm tắt**: Elizabeth phải nhập viện sau khi nhập tiệc với Peter

nêu được tên của bài báo có liên quan (tiêu đề bài báo, tác giả, năm), phương pháp giải quyết của bài báo đó, dữ liệu thực nghiệm, kết quả đạt được và các hạn chế nếu có

Gợi ý: Nên tìm kiếm các bài báo survey để có cái nhìn tổng quát về vấn đề cần giải quyết và các phương pháp giải quyết bài toán

## 2.2 Phương pháp đề xuất giải quyết bài toán

Nhìn một cách tổng quan thì con người luôn linh hoạt hơn trong việc tóm tắt một cái gì đó và di nhiên không thể đi khớp được với phương pháp tóm tắt bằng cách trích chọn như đã trình bày trên mà là hướng tóm lược nhưng đa số các nghiên cứu được thực hiện trong thời gian gần đây đều ứng dụng hướng trích chọn là chủ yếu vì chính hướng trích chọn này cho ra kết quả khả quan và tối ưu hơn mặc dù ta phải hi sinh tính linh hoạt trong tư duy của con người hiện hữu trong mô hình này.

Dù vậy, mô hình tóm tắt bằng hướng tóm lược – abstractive sẽ được áp dụng cho bài làm của chúng em lần này. Với lượng kiến thức còn non trẻ cộng thêm tính căng thẳng của phương pháp tóm tắt bằng tóm lược khi chúng ta phải đương đầu với các vấn đề như cách thể hiện ngữ nghĩa, suy luận và sinh ngôn ngữ tự nhiên và để thực hiện được tốt, chúng ta cần nhiều thời gian hơn để nghiên cứu cũng như triển khai và đánh giá thì chất lượng kết quả trả ra có thể sẽ không được cao như các mô hinh khác đã hiện hữu trên thị trường. Nhưng tóm tắt bằng hướng tóm lược luôn được mong đợi là cách tóm tắt tối ưu và giống con người nhất nên việc thử sức với phương pháp này là một trải nghiệm thú vị.

## 2.3 Phương pháp giải quyết bài toán

### 2.3.1 Mô hình tổng quát

Vẽ được mô hình tổng quát giải quyết bài toán và trình bày tổng quát về từng phần trong mô hình này

Ví dụ:

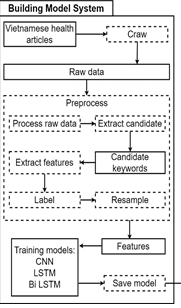
Mô hình tổng quát giải quyết bài toán Keyword Extraction được trình bày theo sơ đồ dưới đây. Trong mô hình này có gồm có 4 phần chính như sau:

Phần 1: Thu thập dữ liệu tự động (crawl)

Phần 2: Trích xuất feature

Phần 3: Trích xuất ứng viên Keyword

Phần 4: Huấn luyện bằng các mô hình học sâu: CNN, LSTM và Bi-LSTM



### 2.3.2 Đặc trưng của mô hình đề xuất

Trình bày cụ thể chi tiết từng thành phần trong mô hình đề xuất trên.

Ví dụ với mô hình ở trên:

2.3.2.1. Cào dữ liệu

Trình bày về cách thu thập và xử lý dữ liệu

2.3.2.2 Trich xuất đặc trưng

Our team will extract features for candidate keywords from five concepts:

* Word Chúng embedding with Pre-trained Word2Vec
* Named-Entity Recognition
* Frequency-based
* Position
* Phrase Length and Word Length

**Word Chúng embedding**

Every word has reflected the structure of the word regarding the schúng emantical/morphological/context/hierarchical/etc. information. The idea of Word Chúng embedding is to capture with thchúng em as much as possible and convert it to vectors. We applied Word Chúng embedding to represent a keyword – can be combined by 2 or more words – as a vector by plus vectors to each word in the dictionary we build.

Ex: Vec “Bệnh viện Trung Ương” = Vec”Bệnh viện”+ Vec”Trung Ương”

Vec “Đại học FPT thành phố Hồ Chí Minh” = Vec “Đại học” + Vec “FPT” + Vec “thành phố” + Vec “Hồ Chí Minh”

**Named-entity recognition**

In the research time, we realized that a noun phrase such as “Tp. Hồ Chí Minh”, “Đại học FPT”, “Bệnh viện Từ Dũ” can be a keyword. A noun phrase containing a named-entity that is considered a keyword is higher than others.

Number of words in Noun Phrase: nNP

Number of named-entity words in Noun Phrase: nNER

We combine both single value and the following formula to have a single feature value (Fner):

For example: Consider the keyword “Đại học FPT thành phố Hồ Chí Minh”.

Words: “Đại học”, “FPT”, “thành phố”, “Hồ Chí Minh”

nNP = 4

nNER = 2 (“FPT”, “Hồ Chí Minh”)

**Frequency-based**

If a noun phrase is occurring more frequently in a document, the phrase is assumed to more important in the document. [3]

Number of times a noun phrase occurs independently in a document: Phrase Frequency (PF).

Number of times a noun phrase appears in full as a part of other noun phrases: Phrase Link Count (PLC)

Two features, PF and PLC, are combined to have a single feature value using the following measure:

In the above formula, the frequency of a noun phrase (PF) is squared only to give it more important than the phrase link count (PLC). The value 1/2 has been used to moderate the value. We explain below about this formula with an example:

Assume that we have 3 keywords and PF and PLC for 2 keywords:

K1 = “sữa mẹ tốt cho trẻ chúng em”, K2 = “sữa mẹ”, K3 = “sữa”

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PF** | **PLC** | **PF+PLC** |
| **K2** | 10 | 5 | 15 |
| **K3** | 1 | 14 | 15 |

So, for these two cases, the simple addition of PF and PLC do not make any difference in assigning weights to the noun phrases although the independent occurrence of noun phrase K2 is more than that of the noun phrase K3. However, the independent existence of a phrase should get higher importance while deciding whether a noun phrase is keyword worthy or not.

For another more general case about PF and PLC with the same K2 and K3:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PF** | **PLC** | **PF+PLC** |
| **K2** | 14 | 1 | 15 |
| **K3** | 1 | 19 | 20 |

In this situation, the simple addition of PF and PLC will favor the K3 “sữa”, but our formula will give a higher score to the K2 “sữa mẹ” because it occurs more independently than the K3 “sữa”.

Inverse document frequency (IDF) is a useful measure to determine the commonness of a term in a corpus.

N = total number of documents in a corpus.

df (document frequency) = the number of documents in which a term occurs.

A term with a lower df value means the term is less frequent in the corpus and hence idf value becomes higher. So, if idf value of a term is higher, the term is relatively rare in the corpus. In this way, idf value is a measure for determining the rarity of a term in a corpus.

In TF-IDF, where TF indicates the frequency of a term in a document, TF\*IDF measure favors a relatively rare term which is more frequent in a document. We combine and IDF in the following formula to have a variant of Edmundsonithe an thchúng ematic feature:

The value of this feature is normalized by dividing the value by the maximum score in a colletion of scores obtained by the phrases corresponding to a document.

**Position**

If a noun phrase occurs in the title or abstract of a document, it should be given more score. So, we consider the position of the first occurrence of a noun phrase in a document as a feature. Unlike the previous approach[3] that assumes the position of a noun phrase occurred first in the sentence I in the document and computed as a feature, in our work, we assume the position of a noun phrase occurred first in the document as I and computed using the following formula from the previous approach.

For example, assume that we have the sentence “Cúm A/H1N1 trở lại tại thành phố Hồ Chí Minh.” and the keyword “Cúm A/H1N1”:

I of “Cúm A/H1N1” = 1:

By this feature, we are easy to predict a keyword in the title of a document where is near to 1 and in summary of a document where is near to 0.

**Phrase Length and Word Length**

These two features can be considered as the structural features of a noun phrase. Noun phrase length becomes an important feature in the keyword extraction task because the length of the noun phrase usually varies from 1 word to 5 words. We find that noun phrase consisting of 6 or more words are relatively rare in our corpus.

Length of the words in a noun phrase can be considered as a feature. According to Zipf’s Law, shorter words occur more frequently than the larger ones. For example, articles occur more frequently in text. So, the word length can be an indication for the rarity of a word. In the previous approach[3] that considered a length of the longest word in noun phrase as a feature, we realized that a maximum length of single words in Vietnamese is 7 and 80% words in Vietnamese is a compound word, that is not the same as English. We decided to choose the length of characters in the noun phrase as a feature.

Length of a noun phrase: PL

Length of characters in the noun phrase: WL

We combine both single value and the following formula to create a feature[3]

For example, assume that we have the keyword “Đại học FPT thành phố Hồ Chí Minh”:

Keyword after tokenize: “Đại\_học FPT thành\_phố Hồ\_Chí\_Minh”

PL = 4, WL = 33

The value of this feature is normalized by dividing the value by the maximum value of the feature in the collection of phrases corresponding to a document.

**2.3.2.3 Trích xuất ứng viên Keyword (Extracting candidate keywords)**

To extract the candidate keywords, we used Part-of-Speech (POS) Tagging and Chunking to extract a noun phrase.

POS Tagging is the process of marking a word in text (corpus) as corresponding to a particular part of speech based on both its definition and its context. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. In Vietnamese, we used Pyvi library from python to pos tagging Vietnamese words.

For example: “Hôm nay trời đẹp quá”

“ N N A R “ with N (noun), A(Adjective), R(Adverb)

Chunking or Noun Phrase Chunking (NP-chunking) is a natural language process that separates and segments a sentence into its sub constituents, such as noun, verb, and noun phrase. One of the most useful sources of information for NP-chunking is part-of-speech tags. In order to create a noun phrase, we first defined a chunk grammar, consisting of rules that indicate how sentences should be chunked. In our problchúng em, we defined a grammar with a single regular-expression rule.

(<N.\*>+ <A>\* <E>)? <N.\*>+ <A>\*

<N.\*>+: One or more of any type nouns

<A>\*: Any number of adjectives

<E>: One or zero of preposition

So, this grammar meaning is a noun phrase can be combined by 2 others noun phrase with preposition word between.

In NP-chunking, we used nltk library from python to chunk a sentence. Using this above grammar and nltk chunking, we create a result tree, from which we can extract a noun phrase.

For example:

“Nguyên là một sinh viên giỏi của trường đại học FPT ở thành phố Hồ Chí Minh”

Calling a noun phrase we need is candidate keyword (CK)

Result tree after chunking:

Figure 30: Example of chunking

S(Sentence), V(Verb), Np(Proper Noun), M(Numeral), Ny(Noun abbreviation)

After all, we get 3 noun phrase: “Nguyên”, “sinh viên giỏi của trường đại học FPT”, “thành phố Hồ Chí Minh”. Sometimes, NP-Chunking by nltk get the noun phrase like “anh ấy”, we call this noun phrase is stop-word and rchúng emove it out of list noun phrase. In the experimental process, we found that noun phrase consisting of 6 or more words are relatively rare in our corpus. So we filter out all noun phrase that has 6 or more word in extracting process.

2.3.2.4 Phương pháp huấn luyện

CNN

Trình bày cụ thể về mô hình CNN và cách thức hoạt động của nó trong bài toán Keyword extraction

LSTM

Trình bày cụ thể về mô hình LSTM và cách thức hoạt động của nó trong bài toán Keyword extraction

Bi-LSTM

Trình bày cụ thể về mô hình Bi-LSTM và cách thức hoạt động của nó trong bài toán Keyword extraction

**……**

# PHẦN III

# THỰC NGHIỆM

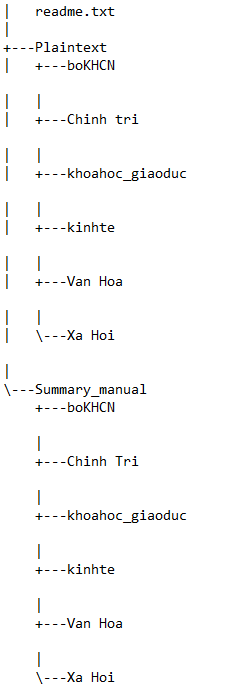
## 3.1 Dữ liệu

### 3.1.1 Nguồn gốc dữ liệu

Dữ liệu – dataset được chúng em lấy trực tiếp từ tác giả [4] Lê Thanh Hương và đã có trích dẫn ở mục Tài liệu tham khảo bên dưới.

### 3.1.2 Mô tả dữ liệu

Dữ liệu mới khi tải về sẽ được đóng gọi dạng file .zip và có cấu trúc thư mục như sau:

**

#### Hình 3.1: Cây thư mục – Folder tree tập dữ liệu

* Folder “Plaintext” là nơi chứa các bài báo nguyên văn được phân chia theo danh mục các chủ đề
* Folder “Summary\_Manual” là nơi chứa các bản tóm tắt thủ công cho từng bài báo tương ứng trong folder “Plaintext” và cũng được phân chia theo danh mục các chủ đề
* File “readme” đặc tả chi tiết dữ liệu từ phía tác giả, nội dung đầy đủ như sau:

*Tập ngữ liệu tóm tắt đơn văn bản:*

*Tập ngữ liệu liệu chứa 200 văn bản và các bản tóm tắt tương ứng, được tổ chức như sau. Tập ngữ liệu gồm có hai thư mục lớn:*

*+ Thư mục Plaintext chứa các văn bản gốc. Các văn bản này là các bản tin được lấy trên các báo điện tử như Dantri, VnExpress, và trên trang web của Bộ Khoa học Công nghệ. Các bản tin này được loại bỏ đoạn tóm tắt in đậm và các liên kết đến các bản tin khác ở đầu bản tin cũng như các hình ảnh trong bản tin. Các bản tin trên các báo Dantri và VnExpress được lấy ở các thể loại Chính trị, Khoa học giáo dục, Kinh tế, Văn hóa, Xã hội. Trong thư mục Plaintext có các thư mục con tương ứng với các thể loại đó.*

*+ Thư mục Summary\_manual chứa các bản tóm tắt. Mỗi bản tóm tắt có cùng tên với văn bản gốc trong thư mục Plaintext. Thư mục này cũng được chia thành các mục thư con tương ứng với các thư mục ở Plaintext.*

*Các bản tóm tắt được tạo ra một cách thủ công. Mỗi văn bản gốc có một bản tóm tắt tương ứng. Các bản tóm tắt có độ dài xấp xỉ 100 từ.*

## 3.2 Xử lý dữ liệu

Có phải tiền xử lý dữ liệu không? Tại sao

Xử lý dữ liệu như thế nào?

## 3.3 Công nghệ sử dụng

(Phần này trình bày không quá 3-6 dòng)

Trình bày vắn tắt về tên của Ngôn ngữ lập trình sử dụng ,

Các thư viện sử dụng để hiện thực bài toán,

Các công cụ sử dụng giải quyết bài toán

* Ngôn ngữ sử dụng: Python 3.9.4
* Word2Vec train: genism
* Ngrams: nltk
* Đọc và ghi file: Pandas

Ví dụ tham khảo:

* Collecting data: Selenium 3.12.0
* Tokenization & POS Tagging: Pyvi 0.0.9.1 (Tran Viet Trung 2016)
* Named-Entity Recognition: underthesea 1.1.8
* Chunking: nltk 3.3
* Word2vector pre-train: streetcodevn (Hung Le 2018)
* Word2vector train: gensim 3.4.0

## 3.4 Cách đánh giá

Cách đánh giá bằng độ đo gì, độ đo đó được tính bằng công thức nào?

## 3.5 Kết quả đạt được

Trình bày các tham số cụ thể hóa các thực nghiệm: ví dụ như số epoch, số tầng, số lớp, tham số x, …

Trình bày code của các module chính

Trình bày các kết quả đạt được theo các độ đo ở trên và trên các bộ dữ liệu hay các phương pháp học máy, học sâu khác nhau.

So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ so sánh giữa các phương pháp theo các độ đo đã trình bày ở trên ví dụ như: Accuracy, MSE, RMSE, MAP, …. (hãy lựa chọn ít nhất 2 độ đo trong các độ đo phổ biến để đánh giá bài toán trên)

Giải thích tại sao đạt được kết quả đó, và kết quả khác nhau giữa các phương pháp trên là nguyên nhân nào?

# PHẦN IV

# KẾT LUẬN

4.1 Kết quả đạt được

Trình bày tóm tắt các kết quả đạt được của bài toán trên,

Trình bày được hạn chế của phương pháp giải quyết bài toán

4.2 Hướng phát triển

Trình bày hướng phát triển trong tương lai

# TÀI LIỆU THAM KHẢO

**Chú ý các tài liệu tham khảo phải được chỉ mục vào trong các Chương nếu có tham khảo, phải có ít nhất từ 7-15 tài liệu tham khảo chính thống: Sách, bài báo Hội nghị, tạp chí - journal, … và tài liệu này phải được đánh chỉ mục từ nhỏ tới lớn và gắn vào trong báo cáo.**

**Các hình ảnh tham khảo hay các phần tham khảo ít thì có thể đánh thành footnote ngay trong trang tham khảo**

1. FloydHub - *“*[*A Gentle introduction to Text Summarization in Machine Learning*](https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/)*”*
2. Nguyễn Viết Hạnh *- Luận văn thạc sĩ công nghệ thông tin về NGHIÊN CỨU TÓM TẮT VĂN BẢN TỰ ĐỘNG VÀ ỨNG DỤNG*
3. Nguyễn Mạnh Cường *– Luận văn thạc sĩ khoa học máy tính về TÓM TẮT VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG DỰA TRÊN MÔ HÌNH ĐỒ THỊ*
4. Github Repo của nhóm làm việc – [*Vietnamese-Text-summarization*](https://github.com/NhutNguyen236/Vietnamese-Text-summarization)
5. Lê Thanh Hương - [*Vietnamese text summarization corpus (kho ngữ liệu tóm tắt văn bản tiếng Việt)*](https://users.soict.hust.edu.vn/huonglt/)
6. Nguyễn Minh Thành – [*Báo cáo về tóm tắt văn bản tự động*](https://sites.google.com/site/trangmonhocitc/text-summarization)
7. Hobson Lane, Cole Howard, Hannes Max Hapke *- Natural Language Processing in action*
8. Maofu Liu, Wenje Li, Mingli Wu, Quin Lu – [*Extractive Summarization based on Event term clustering*](https://www.aclweb.org/anthology/P07-2047.pdf)
9. Trần Mai Vũ – [*Hệ thống tóm tắt sự kiện*](https://eprints.uet.vnu.edu.vn/eprints/id/eprint/3716/1/Event%20Summarization%20System%20%28Technical_report%29.pdf)

# PHỤ LỤC

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

# TỰ ĐÁNH GIÁ

(Với nhóm có 2 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Phần | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| I (1) | Phân tích yêu cầu của bài toán | 1 | 1 |  |
| II (3.25) | Phương pháp giải quyết bài toán | | | |
|  | 2.1 Các phương pháp giải quyết bài toán | 1 | 1 |  |
|  | 2.2 Phương pháp đề xuất giải quyết bài toán | 0.25 | 0.25 |  |
|  | 2.3 Phương pháp giải quyết bài toán | 2 |  |  |
| III (4.75) | Thực nghiệm | | | |
|  | 3.1 Dữ liệu | 0.5 | 0.5 |  |
|  | 3.2 Xử lý dữ liệu | 1 |  |  |
|  | 3.3 Công nghệ sử dụng | 0.25 | 0.25 |  |
|  | 3.4 Cách đánh giá | 0.5 |  |  |
|  | **3.5 Kết quả đạt được** | 2.5 |  |  |
| **IV (0.25)** | **Kết luận** | 0.25 |  |  |
|  | Điểm nhóm | 0.5 | 0.5 |  |
|  | Báo cáo theo mẫu | 0.25 | 0.25 |  |
| Tổng điểm đạt được | |  | |  |