TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN NHẬP MÔN TRÍ TUỆ NHÂN TẠO**

**…tên đề tài…**

*Người hướng dẫn*: **TS NGUYỄN VĂN A**

*Người thực hiện*: **NGUYỄN THỊ B – MSSV**

**TRẦN VĂN C – MSSV**

Lớp **: 10050301**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN ………………….**

**…tên đề tài…**

Người hướng dẫn: **TS NGUYỄN VĂN A**

Người thực hiện: **NGUYỄN THỊ B**

**TRẦN VĂN C**

Lớp **: 10050301**

Khoá  **: 16**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Đây là phần tác giả **tự viết** ngắn gọn, thể hiện sự biết ơn của mình đối với những người đã giúp mình hoàn thành Đồ án. Tuyệt đối không sao chép theo mẫu những “lời cảm ơn” đã có.

**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 Nguyễn Văn A;. 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 tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Văn B*

*Trần Văn C*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Trình bày tóm tắt vấn đề nghiên cứu, các hướng tiếp cận, cách giải quyết vấn đề và một số kết quả đạt được, những phát hiện cơ bản trong vòng 1 -2 trang.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc387692905)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN i](#_Toc387692906)

[TÓM TẮT i](#_Toc387692907)

[MỤC LỤC 1](#_Toc387692908)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 1](#_Toc387692909)

[CHƯƠNG 1 – MỞ ĐẦU 1](#_Toc387692910)

[1.1 Tiểu mục cấp 1 1](#_Toc387692911)

[1.1.1 Tiểu mục cấp 2 1](#_Toc387692912)

[1.1.1.1 Tiểu mục cấp 3 1](#_Toc387692913)

[1.1.1.2 Tiểu mục cấp 3 tiếp theo. 1](#_Toc387692914)

[1.1.2 Tiểu mục cấp 2 tiếp theo 1](#_Toc387692915)

[1.2 Nội dung của chương này 1](#_Toc387692916)

[CHƯƠNG 2 – TỔNG QUAN 1](#_Toc387692917)

[1.1 Trình bày công thức toán học 1](#_Toc387692918)

[1.2 Trình bày một hình vẽ, sơ đồ 1](#_Toc387692919)

[CHƯƠNG 3 – CƠ SỞ LÝ THUYẾT / NGHIÊN CỨU THỰC NGHIỆM 1](#_Toc387692920)

[3.1 Chèn bảng: 1](#_Toc387692921)

[3.2 Viết tắt 1](#_Toc387692922)

[3.3 Trích dẫn 1](#_Toc387692923)

[3.3.1 Tài liệu tham khảo và cách trích dẫn 1](#_Toc387692924)

[3.3.2 Qui định của Khoa Công nghệ thông tin 1](#_Toc387692925)

**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 BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

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

**DANH MỤC BẢNG**

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

PHẦN I

CHƯƠNG 1

PHÂN TÍCH YÊU CẦU BÀI TOÁN

Chương này Phân tích được yêu cầu của bài toán là gì (0.5 điểm)

1.1 Giới thiệu về bài toán

Sử dụng **kiểu chữ Times New Roman (Unicode) cỡ 13** của hệ soạn thảo Winword; **mật độ chữ bình thường**; không được nén hoặc kéo dãn khoảng cách giữa các chữ; **dãn dòng đặt ở chế độ 1.5 lines**; **lề trên 3.5 cm**; **lề dưới 3cm**; **lề trái 3.5 cm**; **lề phải 2cm**. Số trang được đánh ở giữa, phía trên đầu mỗi trang giấy. Nếu có bảng, biểu, hình vẽ trình bày theo chiều ngang khổ giấy thì đầu bảng là lề trái của trang, nhưng nên hạn chế trình bày theo cách này.

Nội dung của tiểu mục cấp 1, một mục khi chia nhỏ thì tối thiểu là 02 mục con (tức là nếu có 1.1.1 thì phải có 1.1.2); tối đa không nên quá 05 mục con.

1.1.1 Tiểu mục cấp 2

Nội dung chi tiết của tiểu mục.

1.1.1.1 Tiểu mục cấp 3

- Đây là cấp tiểu mục nhỏ nhất, không thể tiếp tục phân chia.

- Các ý trong tiểu mục được trình bày gạch đầu dòng “-“.

- Các ý nhỏ hơn sử dụng bullet như sau:

* Ý nhỏ 1.
* Ý nhỏ 2.

- Cần lưu ý rằng đây là cấp sâu nhất, không được phép chia thành 1.1.1.1.1 .

1.1.1.2 Tiểu mục cấp 3 tiếp theo.

Nội dung của tiểu mục thứ ba, khi soạn thảo hãy dùng Styles có sẵn, để khi tạo mục lục sẽ tự động và đồng nhất mỗi khi chúng ta thay đổi format.

1.1.2 Tiểu mục cấp 2 tiếp theo

Không phải lúc nào cũng chia thành tiểu mục cấp 3, nếu như ý trình bày được gói gọn.

1.2 Phân tích yêu cầu của bài toán

Chương này trình bày lý do chọn đề tài, mục đích, đối tượng và phạm vi nghiên cứu, ý nghĩa khoa học và thực tiễn của đề tài; cơ sở khoa học của việc chọn đề tài...;

CHƯƠNG 2

TỔNG QUAN

*Tổng quan*: phân tích, đánh giá các công trình nghiên cứu đã có của tác giả, các tác giả khác trong và ngoài nước liên quan mật thiết đến đề tài; nêu những vấn đề còn tồn tại; chỉ ra những vấn đề mà đề tài cần tập trung nghiên cứu, giải quyết;

1.1 Trình bày công thức toán học

Công thức toán học cần nhất quán trong toàn bộ luận văn. Sử dụng công cụ của MS Word và đánh số theo chương, số thứ tự trong chương. Ví dụ công thức sau đây đánh là 2.1.

(2.1)

Trong công thức, nếu có các kí hiệu là lần đầu tiên sử dụng, cần phải giải thích rõ kí hiệu đó đại diện cho phần tử nào và đơn vị là gì.

(2.2)

Trong công thức 2.2:

* a: độ dài cạnh góc vuông của tam giác, đơn vị mét (*m*).
* b: độ dài cạnh góc vuông còn lại của tam giác, đơn vị mét (*m*).
* c: độ dài cạnh huyền của tam giác, đơn vị mét (*m*).

1.2 Trình bày một hình vẽ, sơ đồ

Việc đánh số bảng biểu, hình vẽ, phương trình phải gắn với số chương (Thí dụ hình 3.4 có nghĩa là hình thứ 4 trong Chương 3). Mọi bảng biểu, đồ thị lấy từ các nguồn khác phải được trích dẫn đầy đủ (ví dụ: *Nguồn:* *Tr.35, Tạp chí Tín dụng (2012), Số 15, NXB Tài chính, TP.HCM*). Nguồn được trích dẫn phải được liệt kê chính xác trong Danh mục tài liệu tham khảo. Đầu đề của bảng, biểu ghi phía trên bảng, biểu; đầu đề của hình vẽ ghi phía dưới hình. Thông thường những bảng ngắn và đồ thị nhỏ phải đi liền với phần nội dung đề cập tới các bảng và đồ thị này ở *lần thứ nhất*. Các bảng dài có thể để ở những trang riêng nhưng cũng phải tiếp ngay theo phần nội dung đề cập tới bảng này ở lần đầu tiên.



Hình 2.1: Kiến trúc FTP

Nguồn: (Postel & Reynolds 1985)

CHƯƠNG 3

CƠ SỞ LÝ THUYẾT / NGHIÊN CỨU THỰC NGHIỆM

*Những nghiên cứu thực nghiệm hoặc lý thuyết*: trình bày các cơ sở lý thuyết, lý luận, giả thuyết khoa học và phương pháp nghiên cứu sẽ được sử dụng trong Luận văn, Luận án;

3.1 Chèn bảng:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tiêu đề A** | **Tiêu đề B** |
| 1 | Nội dung 1 | Nội dung 4 |
| 2 | Nội dung 2 | Nội dung 5 |
| 3 | Nội dung 3 | Nội dung 6 |

Bảng 3.1Ví dụ cho chèn bảng

Khi cần chèn tên bảng thì chọn References \ Caption và chọn “Bảng …”

3.2 Viết tắt

Không lạm dụng việc viết tắt. Chỉ viết tắt những từ, cụm từ hoặc thuật ngữ *được sử dụng nhiều lần trong luận văn*. Không viết tắt những cụm từ dài, những mệnh đề hoặc những cụm từ ít xuất hiện. Nếu cần viết tắt những từ, thuật ngữ, tên các cơ quan, tổ chức... thì được viết tắt sau lần viết thứ nhất có kèm theo chữ viết tắt trong ngoặc đơn. Nếu có quá nhiều chữ viết tắt thì phải có bảng danh mục các chữ viết tắt (xếp theo thứ tự A, B, C) ở phần đầu luận văn.

3.3 Trích dẫn

3.3.1 Tài liệu tham khảo và cách trích dẫn

Mọi ý kiến, khái niệm, phân tích, phát biểu, diễn đạt... có ý nghĩa, mang tính chất gợi ý *không phải của riêng tác giả* và mọi tham khảo khác **phải được trích dẫn và chỉ rõ nguồn trong danh mục Tài liệu tham khảo của luận văn**. Phải nêu rõ cả việc sử dụng những đề xuất hoặc kết quả của đồng tác giả (*đối với công trình đã công bố khác thì phải trích dẫn bình thường như một tài liệu tham khảo*). Nếu sử dụng tài liệu của người khác và của đồng tác giả (bảng biểu, hình vẽ, công thức, đồ thị, phương trình, ý tưởng...) mà không chú dẫn tác giả và nguồn tài liệu thì **luận văn không được duyệt để bảo vệ**.

Không trích dẫn những kiến thức phổ biến, mọi người đều biết tránh làm nặng nề phần tham khảo trích dẫn.

Nếu người dẫn liệu không có điều kiện tiếp cận được một tài liệu gốc mà phải trích dẫn thông qua một tài liệu khác của một tác giả khác, thì phải nêu rõ cách trích dẫn (*lưu ý phải ghi* *đúng nguyên văn từ chính tài liệu tham khảo và hạn chế tối đa hình thức này).* Nếu cần trích dẫn dài hơn thì phải tách phần này thành một đoạn riêng khỏi phần nội dung đang trình bày, in nghiêng, với lề trái lùi vào thêm 2 cm. Khi mở đầu và kết thúc đoạn trích này không phải sử dụng dấu ngoặc kép. Việc trích dẫn là theo thứ tự của tài liệu ở danh mục Tài liệu tham khảo và được đặt trong ngoặc vuông, khi cần có cả số trang, ví dụ [15, tr.314-315]. Đối với phần trích dẫn từ nhiều tài liệu khác nhau, số của từng tài liệu được đặt độc lập trong từng ngoặc vuông, theo thứ tự tăng dần, ví dụ [19], [25], [41], [42].

3.3.2 Qui định của Khoa Công nghệ thông tin

**- Đạo văn** là việc sử dụng từ ngữ hay ý tưởng của người khác như là của mình trong hoạt động học thuật nói riêng và trong hoạt động sáng tạo nói chung. Tại Đại học Hoa Sen, những hành vi sau đây được xem là đạo văn:

* Sao chép nguyên văn **02** (hai) câu liên tiếp mà không dẫn nguồn đúng quy định;
* Sao chép nguyên văn **03** (ba) câu không liên tiếp mà không dẫn nguồn đúng quy định;
* Diễn đạt lại (rephrase) hoặc dịch (translate) toàn bộ một ý nào đó của người khác mà không dẫn nguồn đúng quy định;
* Sử dụng hơn 30% nội dung của một báo cáo cuối kỳ do chính mình viết để nộp cho 2 lớp khác nhau (cùng học kỳ hoặc khác học kỳ) mà không có sự đồng ý của giảng viên;
* Sao chép một phần hoặc toàn bộ bài làm của người khác.

- Khi luận văn, đồ án, bài tập lớn, được chấm điểm, nếu bị phát hiện đạo văn thì ngay lập tức bị điểm 0. Sinh viên sẽ tiếp tục bị xử lý kỷ luật theo các qui định của Nhà trường.

**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 10-15 tài liệu tham khảo chính thống: Sách, bài báo Hội nghị, tạp chí - journal, …**

1. Quách Ngọc Ân (1992), “Nhìn lại hai năm phát triển lúa lai”, *Di tuyền học ứng dụng*, 98(1), tr. 10-16.
2. Bộ nông nghiệp & PTNT (1996), *Báo cáo tổng kết 5 năm (1992-1996) phát triển lúa lai,* Hà Nội.
3. Nguyễn Hữu Đống, Đào Thanh Bằng, Lâm Quang Dụ, Phan Đức Trực (1997), *Đột biến –* *Cơ sở lý luận và ứng dụng,* Nhà xuất bản nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.
4. Nguyễn Thị Gấm (1996), *Phát hiện và đánh giá một số dòng bất dục đực cảm ứng nhiệt* *độ,* Luận văn thạc sĩ khoa học nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.

……….

1. Anderson J.E. (1985), The Relative Inefficiency of Quota, The Cheese Case, *American* *Economic Review*, 75(1), pp. 178-90.
2. Borkakati R. P.,Virmani S. S. (1997), Genetics of thermosensitive genic male sterility in Rice, *Euphytica* 88, pp. 1-7.
3. Boulding K.E. (1955), *Economics Analysis*, Hamish Hamilton, London.
4. Burton G. W. (1988), “Cytoplasmic male-sterility in pearl millet (penni-setum glaucum L.)”, *Agronomic Journal* 50, pp. 230-231.
5. Central Statistical Oraganisation (1995), *Statistical Year Book*, Beijing.
6. FAO (1971), *Agricultural Commodity Projections (1970-1980)*, Vol. II. Rome.
7. Institute of Economics (1988), *Analysis of Expenditure Pattern of Urban Households in* *Vietnam,* Departement pf Economics, Economic Research Report, Hanoi.

**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

**PHẦN NỘI DUNG CỤ THỂ THEO MỤC LỤC SAU**

PHẦN I

CHƯƠNG 1

PHÂN TÍCH BÀI TOÁN

1.1 Giới thiệu về bài toán

Nêu được bài toán cần giải quyết

1.2 Phân tích yêu cầu của bài toán

Trình bày được các yêu cầu của bài toàn

1.3 Hướng giải quyết

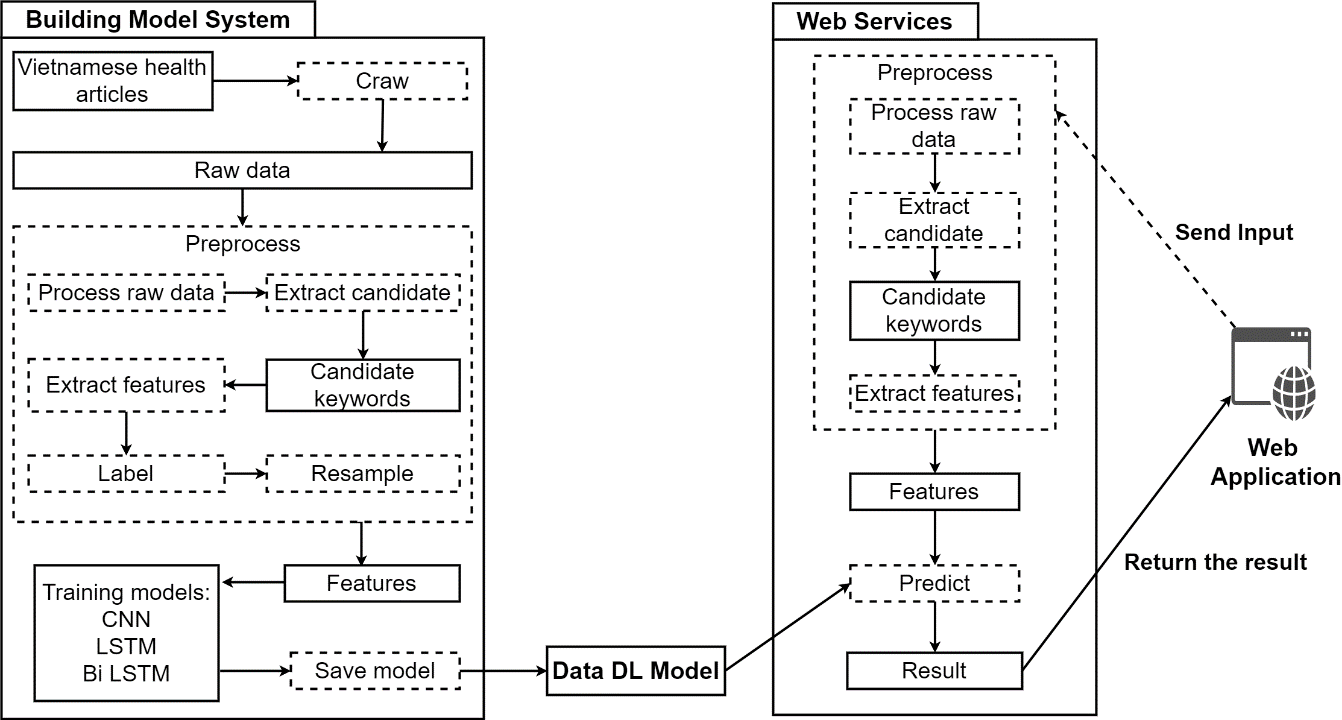
Đưa ra hướng tổng quát giải quyết bài toán

CHƯƠNG 2

2.1 Giới thiệu chung về các phương pháp giải quyết bài toán

Trong phần này phải vẽ được sơ đồ tổng quát giải quyết bài toán

Ví dụ: Mô hình tổng quát giải quyết bài toán Keyword Extraction



2.2 Đặc trưng của mô hình đề xuất

2.2.1 Feature

Ví dụ tham khảo của mô hình trên

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

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

**Word Embedding**

Every word has reflected the structure of the word regarding the semantical/morphological/context/hierarchical/etc. information. The idea of Word Embedding is to capture with them as much as possible and convert it to vectors. We applied Word 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 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 thematic 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.

**2.2.2 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 problem, 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 remove 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.

2.2 Phương pháp 1

CNN

Trình bày cụ thể về giải thuật này

2.3 Phương pháp 2

LSTM

Trình bày cụ thể về giải thuật này

2.4 Phương pháp 3 (đây là phương pháp lựa chọn giải quyết bài toán)

Bi-LSTM

Mỗi phương pháp đều trình bày được sơ đồ giải thuật

Trình bày được ưu nhược điểm của mỗi phương pháp

Với phương pháp 3, giải thích lý do tại sao lựa chọn phương pháp này

CHƯƠNG 3

THỰC NGHIỆM

3.1 Dữ liệu

Dữ liệu được lấy ở đâu, từ nguồn nào

Dữ liệu gồm có những gì

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

Trình bày về Ngôn ngữ lập trình,

Các thư viện,

Các công cụ

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

We evaluate our proposed model base on Precision, Recall and F1-score with test datasets.

All the parameter can be calculated by using leftmost four parameters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predict class** | | |
| **Actual class** |  | *Class = Yes* | *Class = No* |
| *Class = Yes* | True Positive | False Positive |
| *Class = No* | False Negative | True Negative |

*True Positives (TP)* - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes.

*True Negatives (TN)* - These are the correctly predicted negative values which mean that the value of actual class is no and value of predicted class is also no.

*False Positives (FP)* – When actual class is no and predicted class is yes.

*False Negatives (FN)* – When actual class is yes but predicted class in no.

* **Precision**

Precision is the ratio of correctly predicted positive values to the total predicted positive values. This metric highlights the correct positive predictions out of all the positive predictions. High precision indicates low false positive rate.

* **Recall**

The recall is the ratio of correctly predicted positive values to the actual positive values. Recall highlights the sensitivity of the algorithm, i.e. out of all the actual positives how many were caught by the program.

* **F1- Score**

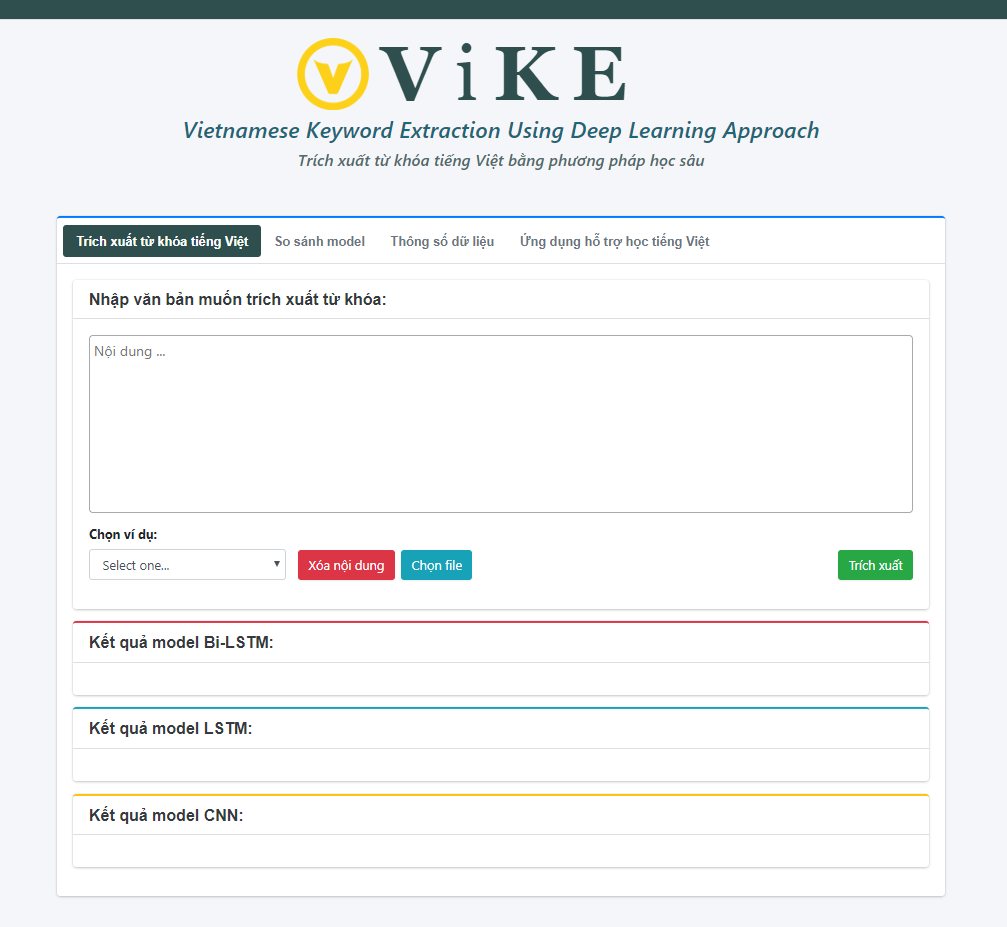
The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

3.5 Kết quả đạt được

Trình bày các kết quả đạt được

Cụ thể hóa các thực nghiệm: tham số, lớp, … và các 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)



CHƯƠNG 4

KẾT LUẬN

4.1 Kết luận

Trình bày tóm tắt các kết quả đạt được

Trình bày được hạn chế của đề tài

4.2 Hướng phát triển

Trình bày hướng phát triển trong tương lai

TỰ CHẤM ĐIỂM

|  |  |  |  |
| --- | --- | --- | --- |
| Điểm Đồ án | | | |
| Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| Chương 1 | 1.5 đ |  |  |
| Chương 2 | 3.0 đ |  |  |
| Chương 3 | 4.0 đ |  |  |
| Chương 4 | 0.5 đ |  |  |
| Tổng điểm | |  |  |
| Điểm nhóm | | | |
| Điểm nhóm | 1đ |  |  |
| Tổng điểm | |  |  |