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**

**Phân lớp văn bản bằng Deep Learning**

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

*Người thực hiện*: **LÂM KIM PHÚ – 51600065**

**TRẦN THỊ MỸ NHI – 518H0543**

Nhóm **: 22**

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**

**Phân lớp văn bản bằng Deep Learning**

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

Người thực hiện: **LÂM KIM PHÚ – 51600065**

**TRẦN THỊ MỸ NHI – 518H0543**

Nhóm **: 22**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Nhóm chúng em xin chân thành cảm ơn những lời chỉ tận tình và sâu sắc nhất của thầy Bùi Thanh Hùng. Qua bài báo cáo tiểu luận này, tụi em đúc kết từ những gì thầy dạy cho tụi em qua các buổi học, và qua những góp ý của thầy.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của 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 24 tháng 04 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lâm Kin Phú*

*Trần Thị Mỹ Nhi*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

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

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 của Bài toán

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 CHỮ VIẾT TẮT 1](#_Toc387692909)

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

PHẦN I

[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)

PHẦN II

[2.1 Tiểu mục 1 1](#_Toc387692918)

[2.2 Tiểu mục 2 1](#_Toc387692919)

PHẦN III

[3.1 Tiểu mục 1 1](#_Toc387692918)

[3.2 Tiểu mục 2 1](#_Toc387692919)

PHẦN IV

[4.1 Tiểu mục 1 1](#_Toc387692918)

[4.2 Tiểu mục 2 1](#_Toc387692919)

PHỤ LỤC

TÀI LIỆU THAM KHẢO

TỰ ĐÁNH GIÁ 6

**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

Text classification sử dụng Long Short Term Memory Networks Deep Learning.

Xây dựng mô hình phân loại văn bản tin tức tiếng Việt cho trang báo điện tử VNExpress. Mỗi khi một bài báo được đăng, chương trình cần phải tự động xác định được bài báo đó nằm trong danh mục nào. Các danh mục gồm có: Kinh tế, Thể thao, Giáo dục, Sức khỏe, Du lịch, Pháp luật,…

1.2 Phân tích yêu cầu của bài toán

Yêu cầu bài toán bao gồm:

* Crawl dữ liệu tự động từ trang báo VNExpress
* Tiền xử lý dữ liệu
* Đưa ra dữ liệu được gắn nhãn tự động và độ chính xác

PHẦN II

PHƯƠNG PHÁP GIẢI QUYẾT

2.1 Các phương pháp giải quyết bài toán

Bài toán này có thể được giải quyết bằng các phương pháp nào, các phương pháp này như thế nào? Trình bày được cụ thể các phương pháp giải quyết bài toán, mỗi phương pháp giải quyết bài toán đều phải 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

Đưa ra nhận định tổng quát về hướng giải quyết bài toán và lý giải tại sao lựa chọn hướng giải quyết này.

2.3 Phương pháp giải quyết bài toán

2.3.1 Mô hình tổng quát

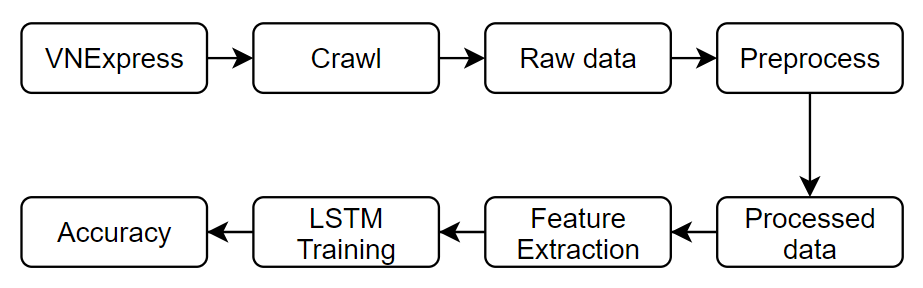
Mô hình tổng quát giải quyết bài toán Text Classification được trình bày theo sơ đồ dưới đây. Trong mô hình này có gồm có 4 phần chính như sau:

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

Phần 2: Tiền xử lý dữ liệu

Phần 3: Trích xuất thuộc tính bằng CountVectorizer

Phần 4: Huấn luyện bằng mô hình học sâu LSTM



2.3.2 Đặc trưng của mô hình đề xuất

2.3.2.1. Cào dữ liệu

* Dữ liệu được lấy từ trang https://vnexpress.net/ . Đầu tiên truy cập vào trang <https://vnexpress.net/> rồi lấy url của các topic (Ví dụ: <https://vnexpress.net/thoi-su>) từ navigation bar của trang. Sau đó vào trang từng của topic để lấy dữ liệu về bằng thẻ <p class="description">. Dữ liệu bao gồm “Topic” chứa topic của bài báo, “Title” chứa tiêu đề bài báo, “Description” chứa phần mô tả sơ lược bài báo.
* Ở phần “Description” có một số bài có gắn location stamp vì thế có một bước để loại bỏ location stamp. Nếu không bỏ thì phần “Description” chữ ở location stamp và phần text nó sẽ bị dính vào nhau.
* Dữ liệu được lấy về có dung lượng khoảng 100KB và được đặt tên là “raw\_data.csv”, bao gồm 3 cột “Topic”, “Title”, “Description”, và khoảng 430 – 460 dòng dữ liệu tùy theo thời điểm crawl.

2.3.2.2 Trich xuất đặc trưng

2.3.2.4 Phương pháp huấn luyện

LSTM – Long Short Term Memory networks

Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng *tanh*.

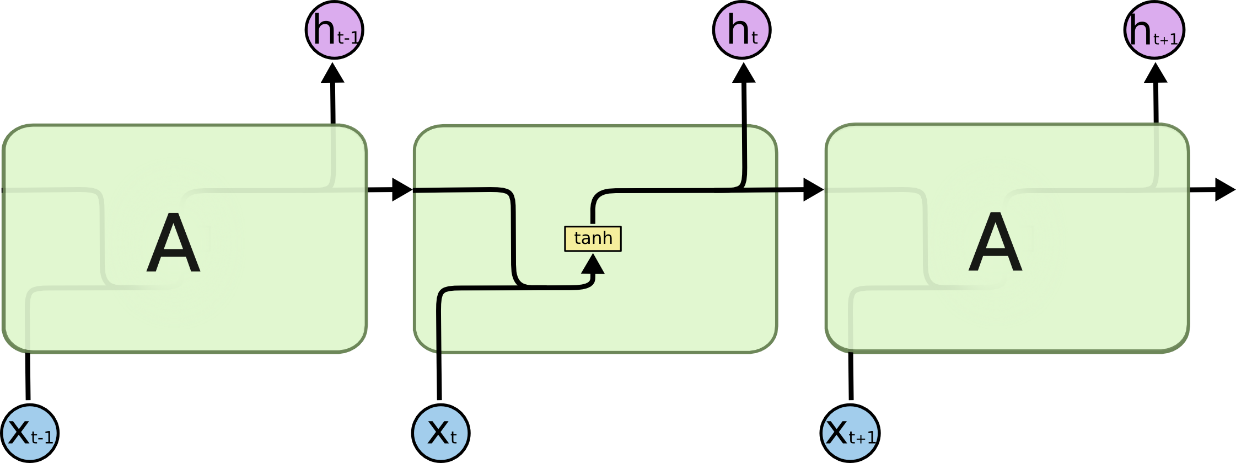


Figure 2.1: The repeating module in a standard RNN contains a single layer.

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

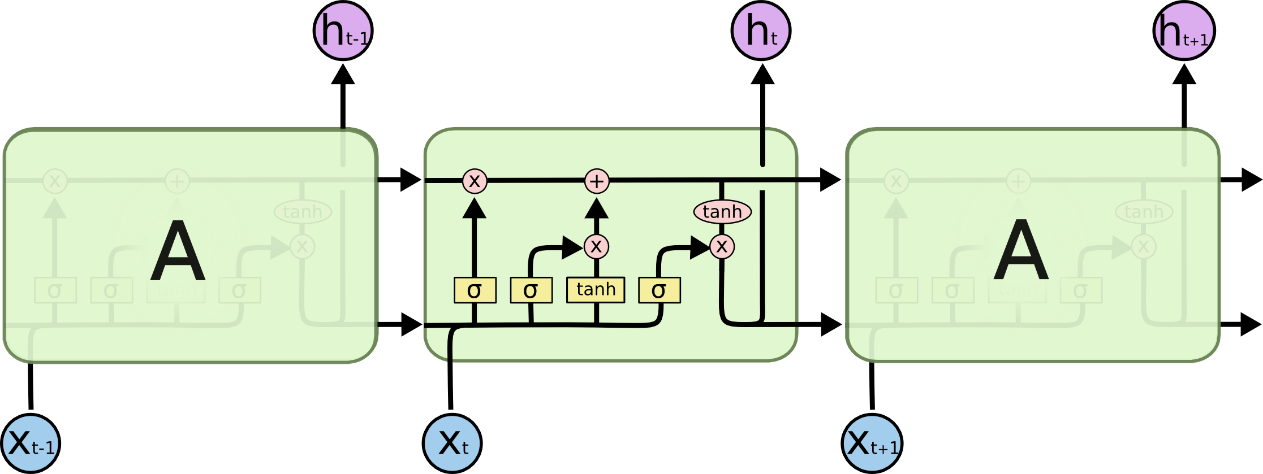


Figure Error! No text of specified style in document.:2The repeating module in an LSTM contains four interacting layers.

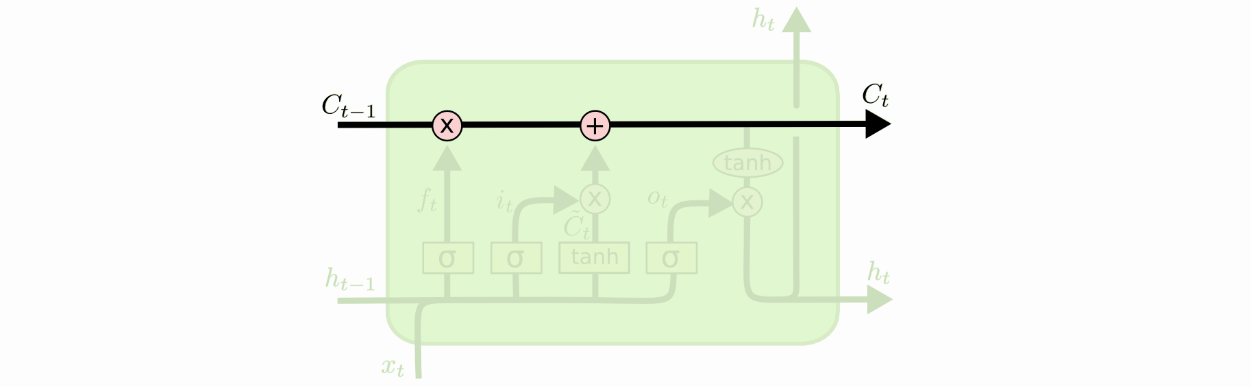
Ở sơ đồ trên, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác. Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau.



Ý tưởng cốt lõi của LSTM

Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính đường chạy thông ngang phía trên của sơ đồ hình vẽ.

Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.



LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho trạng thái tế báo, chúng được điều chỉnh cẩn thận bởi các nhóm được gọi là cổng (gate).

Các cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân.



Tầng sigmoid sẽ cho đầu ra là một số trong khoản [0, 1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 0 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 1 thì có nghĩa là cho tất cả các thông tin đi qua nó.

Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.

PHẦN III

THỰC NGHIỆM

3.1 Dữ liệu

* Dữ liệu được lấy từ trang https://vnexpress.net/ . Đầu tiên truy cập vào trang <https://vnexpress.net/> rồi lấy url của các topic (Ví dụ: <https://vnexpress.net/thoi-su>) từ navigation bar của trang. Sau đó vào trang từng của topic để lấy dữ liệu về bằng thẻ <p class="description">. Dữ liệu bao gồm “Topic” chứa topic của bài báo, “Title” chứa tiêu đề bài báo, “Description” chứa phần mô tả sơ lược bài báo.
* Ở phần “Description” có một số bài có gắn location stamp vì thế có một bước để loại bỏ location stamp. Nếu không bỏ thì phần “Description” chữ ở location stamp và phần text nó sẽ bị dính vào nhau.
* Dữ liệu được lấy về có dung lượng khoảng 100KB và được đặt tên là “raw\_data.csv”, bao gồm 3 cột “Topic”, “Title”, “Description”, và khoảng 430 – 460 dòng dữ liệu tùy theo thời điểm crawl.

3.2 Xử lý dữ liệu

Tiền xử lý dữ liệu là một bước rất quan trọng trong việc giải quyết bất kỳ vấn đề nào trong lĩnh vực Học Máy. Hầu hết các bộ dữ liệu được sử dụng trong các vấn đề liên quan đến Học Máy cần được xử lý, làm sạch và biến đổi trước khi một thuật toán Học Máy có thể được huấn luyện trên những bộ dữ liệu này.

Các bước xử lý dữ liệu:

* Bước 1: lowercase tất cả dữ liệu.
* Bước 2: loại bỏ dấu câu và các kí tự đặc biệt trong dữ liệu.
* Bước 3: tokenize dữ liệu theo thư viện underthesea.
* Bước 4: loại bỏ stopword.

Dữ liệu qua xử lý được lưu vào file “preprocessing.csv” trong thư mục “data”.

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

Ví dụ tham khảo:

* Collecting data: Beautifulsoup 4
* Tokenization: underthesea
* 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. ádas

**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 |  |  |
|  | 2.2 Phương pháp đề xuất giải quyết bài toán | 0.25 | 0.25 |  |
|  | 2.3 Phương pháp giải quyết bài toán | 2 | 2 |  |
| III (4.75) | Thực nghiệm | | | |
|  | 3.1 Dữ liệu | 0.5 | 0.5 |  |
|  | 3.2 Xử lý dữ liệu | 1 | 1 |  |
|  | 3.3 Công nghệ sử dụng | 0.25 | 1 |  |
|  | 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 | | 6.5 | |  |