

**ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN
THÔNG**

=====

NGÔ THANH HẢO

**TÌM HIỂU PHƯƠNG PHÁP PHÂN LOẠI NAÏVE BAYES
VÀ NGHIÊN CỨU XÂY DỰNG ỨNG DỤNG TÓM TẮT
VĂN BẢN TIẾNG VIỆT**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc TS Nguyễn Thị Thu Hà, người đã chỉ bảo và hướng dẫn tận tình cho tôi và đóng góp ý kiến quý báu trong suốt quá trình học tập, nghiên cứu và thực hiện luận văn này.

Tôi xin trân trọng cảm ơn Ban giám hiệu Trường Đại học Công Nghệ Thông Tin và Truyền Thông Đại học Thái Nguyên, khoa CNTT đã giúp đỡ và tạo các điều kiện cho chúng tôi được học tập và làm khóa luận một cách thuận lợi.

Và cuối cùng tôi xin gửi lời cảm ơn đến gia đình, người thân và bạn bè – những người luôn bên tôi và là chỗ dựa giúp cho tôi vượt qua những khó khăn nhất. Họ luôn động viên tôi khuyến khích và giúp đỡ tôi trong cuộc sống và công việc cho tôi quyết tâm hoàn thành luận văn này.

Tuy nhiên do thời gian có hạn, mặc dù đã nỗ lực cố gắng hết mình nhưng chắc rằng luận văn khó tránh khỏi những thiếu sót. Rất mong được sự chỉ bảo, góp ý tận tình của Quý thầy cô và các bạn.

Tôi xin chân thành cảm ơn!

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là kết quả nghiên cứu của tôi, không sao chép của ai. Nội dung luận văn có tham khảo và sử dụng các tài liệu liên quan, các thông tin trong tài liệu được đăng tải trên các tạp chí và các trang website theo danh mục tài liệu của luận văn.

Tác giả luận văn

Ngô Thanh Hảo

MỤC LỤC

LỜI CẢM ƠN	I
LỜI CAM ĐOAN	III
MỤC LỤC	IV
DANH MỤC HÌNH VẼ	VI
DANH MỤC BẢNG BIỂU	VI
DANH MỤC TỪ VIẾT TẮT.....	VIII
LỜI MỞ ĐẦU	1
CHƯƠNG 1 : TỔNG QUAN VỀ TÓM TẮT VÀ TÓM TẮT VĂN BẢN TIẾNG VIỆT	3
1.1 Giới thiệu.....	3
1.1.1 Tổng quan bài toán tóm tắt văn bản	3
1.1.2 Tỷ lệ trong tóm tắt văn bản	6
1.2 Đặc điểm ngôn ngữ tiếng Việt.....	7
1.2.1 Đặc điểm ngữ âm	7
1.2.2 Đặc điểm từ vựng.....	8
1.2.3 Đặc điểm ngữ pháp	9
1.2.4 Xử lý ngôn ngữ tiếng Việt trên máy tính	10
1.3 Một số phương pháp tóm tắt văn bản	12
1.4 Đánh giá tóm tắt văn bản	14
1.4.1 Đánh giá theo cách thủ công	14
1.4.2 Phương pháp đánh giá BLEU	14
1.4.3 Phương pháp đánh giá ROUGE.....	15
1.4.4 Độ đo precision và độ đo recall	16
CHƯƠNG 2 : PHƯƠNG PHÁP TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN NAIVE BAYES	18
2.1 Một số phương pháp tóm tắt văn bản điển hình	18
2.1.1 Phương pháp tóm tắt văn bản bằng cây quyết định	18
2.1.2 Phương pháp tóm tắt văn bản bằng mạng nơ ron	19
2.1.3 Phương pháp phân tích ngôn ngữ tự nhiên mức sâu.....	19
2.1.4 Phương pháp tóm tắt ngắn	22

2.1.5 Phương pháp dựa trên mô hình markov ẩn	23
2.1.6 Phương pháp tóm tắt dựa trên rút gọn câu	24
2.1.7 Phương pháp tóm tắt văn bản bằng naïve bayes:	24
2.2 Phương pháp tóm tắt văn bản sử dụng lý thuyết phân loại Naïve Bayes	25
2.2.1 Phân loại Naïve Bayes	25
2.2.2 Lựa chọn các đặc trưng cho trích chọn	31
2.3 Huấn luyện và tính trọng số các câu trong tập huấn luyện.....	39
2.4 Lựa chọn các câu tạo tóm tắt.....	41
CHƯƠNG 3. XÂY DỰNG VÀ CÀI ĐẶT HỆ THỐNG TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN LÝ THUYẾT NAÏVE BAYES	44
3.1 Mô hình hệ thống tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes	44
3.2 Phân tích thiết kế hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes	50
3.3 Một số giao diện của hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes	52
3.3.1 Giao diện trang chủ hệ thống tóm tắt văn bản tiếng Việt	52
3.3.2 Giao diện trang quản trị hệ thống tóm tắt văn bản tiếng Việt.....	53
3.4 Kết quả thực nghiệm phương pháp tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes.....	59
3.4.1 Xây dựng tập dữ liệu phục vụ huấn luyện	59
3.4.2 Xây dựng bộ từ điển danh từ.....	60
3.4.3 Tiền xử lý và chuẩn hóa dữ liệu.....	60
3.4.4 Đánh giá kết quả của hệ thống tóm tắt văn bản dựa trên Naïve Bayes .	61
KẾT LUẬN	62
TÀI LIỆU THAM KHẢO	63
TIẾNG VIỆT	63
PHỤ LỤC	64

DANH MỤC HÌNH VẼ

Hình 1.1 Hệ Thống Tóm Tắt Văn Bản Text Compactor	4
Hình 2.1. Cây Cấu Trúc Tu Từ	22
Hình 2.2. Mô Hình Markov Ẩn Sử Dụng Trong Trích Rút Câu.	23
Hình 2.3. Ma Trận Ví Dụ.	33
Hình 2.4. Mô Hình Giảm Chiều Véc Tơ.....	33
Hình 2.5. Văn Bản Ví Dụ.....	35
Hình 2.6 Quan Hệ Giữa Số Văn Bản Và Số Thuật Ngữ.....	36
Hình 2.7 Tách Từ Dựa Trên Hệ Thống Phân Tích Câu Visp.....	36
Hình 2.8. Thuật Toán Tinh Trọng Số Của Câu.....	40
Hình 2.9 Thuật Toán Trích Rút Câu	42
Hình 3.1. Mô Hình Tóm Tắt Văn Bản Thông Thường.....	45
Hình 3.2. Mô Hình Tóm Tắt Văn Bản Trong Luận Văn Đề Xuất.....	47
Hình 3.3 Cơ sở dữ liệu của hệ thống.....	50
Hình 3.4 Sơ Đồ Usecase Tổng Quát.	51
Hình 3.5. Usecase Trường Hợp Huấn Luyện.....	52
Hình 3.6. Giao Diện Trang Chủ Của Hệ Thống	53
Hình 3.7 Giao Diện Chính Của Trang Quản Trị.....	54
Hình 3.8 Lấy Tin Tự Động.	54
Hình 3.9 Giao Diện Hiển Thị Dữ Liệu Lấy Về.	55
Hình 3.10 Giao Diện Huấn Luyện Văn Bản.	56
Hình 3.11 Giao Diện Quản Lý Từ.	56
Hình 3.12 Hiển Thị Tin Tức Sau Khi Cập Nhật.	57
Hình 3.13 Giao Diện Tóm Tắt Tin Tức.	58
Hình 3.14 Giao Diện Tóm Tắt Văn Bản.....	58

DANH MỤC BẢNG BIỂU

Bảng 1.1. Hiện Trạng Các Kho Ngữ Liệu Tiếng Việt.....	12
Bảng 2.1 : Ví dụ về bảng huấn luyện.....	28
Bảng 3.1. Bảng Kết Quả Thực Nghiệm.....	61

DANH MỤC TỪ VIẾT TẮT

Kí hiệu	Diễn giải
tf	Tần suất từ (<i>Term frequency</i>)
Idf	tần suất nghịch đảo văn bản (<i>inverse document frequency</i>)
TREC	Hội thảo tra cứu văn bản (<i>Text REtrieval Conference</i>)
DUC	Hội thảo hiểu văn bản (<i>Document Understanding Conference</i>)
BLEU	Phương pháp đánh giá dịch máy tự động (<i>Bilingual Evaluation Under Study</i>)
NIST	Viện công nghệ tiêu chuẩn quốc gia (<i>National Institute of Standards and Technology</i>)
Rouge	Phương pháp đánh giá kết quả tóm tắt ROUGE (<i>Recall – Oriented Understudy for Gisting Evaluation</i>)

LỜI MỞ ĐẦU

Ngày nay thông tin đã và đang đóng vai trò cực kỳ quan trọng trong xã hội. Sự phát triển mạnh mẽ của Internet mang lại cho con người những thông tin quan trọng và bổ ích, với lượng lớn thông tin này mang lại cho con người những tiện ích tra cứu thông tin. Các hệ thống tìm kiếm, tra cứu được nghiên cứu, đề xuất và xây dựng thỏa mãn phần nào yêu cầu của người dùng đặt ra trong hiện tại. Tuy nhiên, nó khiến chúng ta khó khăn trong việc tìm kiếm và tổng hợp thông tin.

Các nhà nghiên cứu đã đề xuất các giải pháp để xây dựng các hệ thống, công cụ khai phá dữ liệu như: phân loại dữ liệu, phân cụm dữ liệu, nén dữ liệu, tra cứu thông tin, tóm tắt văn bản... Một trong những công cụ quan trọng đó là tóm tắt văn bản.

Đối với dữ liệu dạng văn bản, tóm tắt văn bản là tóm tắt các thông tin chính từ trong văn bản gốc để nhận được một văn bản ở dạng ngắn hơn và chất lọc các thông tin quan trọng từ trong văn bản gốc.

Tóm tắt văn bản nhận được nhiều sự quan tâm nghiên cứu của các nhà khoa học nhóm nghiên cứu và các công ty trên thế giới. Bài toán tóm tắt văn bản tiếng Việt cũng không ngoại lệ vì không thể khai thác thông tin tiếng Việt hiệu quả nếu không có phương pháp tóm tắt văn bản tiếng Việt.

Trong khuôn khổ đề tài luận văn, tôi sử dụng cách tiếp cận rút gọn câu dựa trên Naive Bayes để:

- Nâng cao chất lượng của hệ thống tóm tắt văn bản tiếng Việt tự động bằng cách học giám sát. Trên thực tế để giải quyết bài toán này đã có rất nhiều phương pháp được đưa ra như sử dụng thuật toán Naive Bayes, phương

pháp cây quyết định(Decision tree), Phương pháp tóm tắt văn bản bằng mạng nơron nhân tạo(Artificial Neural Network), phương pháp tóm tắt ngắn, Phương pháp phân tích ngôn ngữ tự nhiên mức sâu, phương pháp học không giám sát, phương pháp máy học. Mỗi phương pháp đều cho kết quả khá tốt, tuy nhiên phương pháp tóm tắt văn bản tiếng Việt bằng thuật toán Naïve Bayes có chất lượng của tóm tắt văn bản là cao hơn.

- Giảm độ phức tạp tính toán về mặt thời gian.
- Xây dựng hệ thống tự động tổng hợp tin tức trực tuyến và tóm tắt.
- Xây dựng tập dữ liệu huấn luyện gồm 200 văn bản tiếng Việt.

Luận văn được chia thành 3 chương với các nội dung sau:

Chương 1: Tổng quan về tóm tắt và tóm tắt văn bản tiếng Việt

Chương 2: Phương pháp tóm tắt văn bản tiếng việt dựa trên Naive Bayes

Chương 3: Xây dựng ứng dụng tóm tắt văn bản tiếng Việt dựa trên Naive Bayes.