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

Số hóa bởi Trung tâm Học liệu - ĐHTN http://www.lrc-tnu.edu.vn/

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

Số hóa bởi Trung tâm Học liệu - ĐHTN http://www.lrc-tnu.edu.vn/

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

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

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 chắt 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 Naïve 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.