ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

---------📖---------

NGUYỄN THỊ THÙY DƯƠNG

NGHIÊN CỨU LÝ THUYẾT NAIVE BAYES VÀ ỨNG DỤNG TRONG PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60.48.0101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học:

TS. NGUYỄN THỊ THU HÀ

THÁI NGUYÊN, NĂM 2015

LỜI CAM ĐOAN

Tôi xin cam đoan:

1. Những nội dung trong luận văn này là do tôi thực hiện dưới sự trực tiếp hướng dẫn của cô giáo TS. Nguyễn Thị Thu Hà.
2. Mọi tham khảo dùng trong luận văn đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố.
3. Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo, hay gian trá, tôi xin chịu hoàn toàn trách nhiệm.

Tác giả luận văn

Nguyễn Thị Thùy Dương

LỜI CẢM ƠN

Lời đầu tiên tôi xin được bày tỏ lòng biết ơn chân thành đến Ban Giám Hiệu, các thầy giáo, cô giáo phòng Sau đại học trường Đại học Công Nghệ Thông Tin & Truyền Thông, các thầy giáo ở Viện Công Nghệ Thông Tin đã giảng dạy và tạo mọi điều kiện cho tôi học tập, nghiên cứu và hoàn thành luận văn này.

Đặc biệt, tôi xin bày tỏ sự kính trọng và lòng biết ơn sâu sắc đến TS. Nguyễn Thị Thu Hà, người đã tận tình hướng dẫn và giúp đỡ tôi trong suốt quá trình học tập, nghiên cứu và hoàn thành luận văn.

Tôi chân thành cảm ơn các thầy cô Khoa Công nghệ thông tin, Trường Trung cấp nghề Phát Thanh Truyền Hình Thanh Hóa nơi tôi công tác đã tạo điều kiện và hỗ trợ tôi trong suốt thời gian qua.

Tôi cũng xin chân thành cảm ơn người thân, bạn bè đã giúp đỡ và động viên tôi trong suốt thời gian học tập cũng như trong thời gian thực hiện luận văn.

Xin chân thành cảm ơn!

Thái Nguyên, ngày 20 tháng 08 năm 2015

MỤC LỤC

DANH SÁCH CÁC BẢNG

DANH SÁCH HÌNH VẼ

DANH SÁCH CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
k- NN	k- Nearest Neighbor	k-Láng giềng gần nhất
SVM	Support Vector Machine	Máy véc tơ hỗ trợ
RSS	Really Simple Syndication	Định dạng tập tin
ML	Machine Languages	Ngôn ngữ máy
LSI	Latent Sematic Indexing	Chỉ số ngữ nghĩa ẩn
SVD	Singular Value Decomposition	Phân tích giá trị đơn

MỞ ĐẦU

1. Lý do chọn đề tài

Với lượng thông tin đồ sộ, một yêu cầu lớn đặt ra đối với chúng ta là làm sao tổ chức và tìm kiếm thông tin có hiệu quả nhất. Phân loại thông tin là một trong những giải pháp hợp lý cho yêu cầu trên. Nhưng một thực tế là khối lượng thông tin quá lớn, việc phân loại dữ liệu thủ công là điều không tưởng. Hướng giải quyết là một chương trình máy tính tự động phân loại các thông tin trên.

Đề tài "*Nghiên cứu lý thuyết Naive Bayes và ứng dụng trong phân loại văn bản Tiếng Việt*" nhằm tìm hiểu và thử nghiệm các phương pháp phân loại văn bản áp dụng trên tiếng Việt. Phân loại văn bản (Text classification) là một trong những công cụ khai phá dữ liệu dạng văn bản một cách hữu hiệu, làm nhiệm vụ đưa những văn bản có cùng nội dung chủ đề giống nhau về cùng một lớp có sẵn. Phân loại văn bản giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin cần thiết đồng thời có thể lưu trữ các thông tin theo đúng chủ đề (topic) hay lớp (class) dựa trên các thuật toán phân loại.

2. Đối tượng và phạm vi nghiên cứu: Tìm hiểu lý thuyết Naive Bayes và ứng dụng trong phân loại văn bản tiếng Việt.

3. Những nội dung nghiên cứu chính

- Chương 1: Tổng quan về phân loại văn bản

Tổng quan về phân loại văn bản và khái niệm cơ bản về lý thuyết Naive Bayes, bộ phân loại Naive Bayes trên mô hình xác suất.

- Chương 2: Phân loại văn bản tiếng Việt dựa trên phương pháp Naive Bayes

Trình bày phương pháp phân loại văn bản tiếng Việt dựa trên phân loại Naive Bayes và cách giảm chiều đặc trưng nhằm tăng tốc trong quá trình tính toán xử lý bằng cách sử dụng mô hình chủ đề dùng cho tiếng Việt.