ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

ĐỀ CƯƠNG CHI TIẾT

TÊN ĐỀ TÀI: XÂY DỰNG HỆ THỐNG LỌC RÁC CHO HỆ THỐNG PHÁT HIỆN TIN NÓNG TỪ CÁC TRANG TIN TỨC

Cán bộ hướng dẫn: TS. Huỳnh Ngọc Tín

Thời gian thực hiện: Từ ngày 04/09/2017 đến ngày 04/01/2018

Sinh viên thực hiện:

Hoàng Anh Minh – 13520505

Lâm Tuấn Anh – 13520020

Nội dung đề tài:

1. Mục tiêu đề tài:
   - Tìm và chọn được phương pháp phù hợp nhất để nhận biết tin nóng, là tin về những sự kiện mới, có khả năng thu hút sự chú ý, quan tâm của nhiều người.
   - Xây dựng hệ thống áp dụng phương pháp trên để hỗ trợ biên tập viên trong việc viết bài.
2. Phạm vi đề tài:
   - Nguồn dữ liệu: các bài viết từ báo chính thống Việt Nam.
   - Ngôn ngữ: tiếng Việt.
   - Các thuật toán tìm hiểu và áp dụng: Naive Bayes, SVM, J48
3. Phương pháp thực hiện:
   - Tìm hiểu bài toán phân loại tin tức, tìm hiểu các phương pháp và các hướng tiếp cận
   - Thử nghiệm đánh giá các phương pháp đã tìm hiểu:

     Thu thập dữ liệu tin tức từ cơ sở dữ liệu của công ty VCCorp.

     Tiến hành một số thống kê trên dữ liệu thu thập được.

     Huấn luyện và so sánh kết quả các thuật toán phân lớp: Naive Bayes, SVM, J48.

4. Kết quả mong đợi:
    - Hệ thống lọc rác tin tức
    - Kiến thức về bài toán phân lớp văn bản
    - Báo cáo đề tài

Kế hoạch thực hiện:

- Đặt vấn đề và xây dựng phát biểu bài toán.
- Lên kế hoạch thực hiện đề tài.
- Tìm hiểu bài toán phân lớp văn bản.
- Tìm hiểu bài toán lọc rác.
- Tìm hiểu framework Struts2, thư viện React, Apache Lucene, Weka.
- Thu thập và gán nhãn dữ liệu.
- Xử lý dữ liệu và huấn luyện các mô hình phân lớp.
- Cài đặt và hiện thực hệ thống lọc rác.
- Xây dựng API quản lý dữ liệu tin tức.
- Viết báo cáo.

Xác nhận của CBHD (Ký tên và ghi rõ họ tên)	**TP. HCM, ngày….tháng …..năm…..** **Sinh viên** (Ký tên và ghi rõ họ tên)