**Học viện công nghệ bưu chính viên thông**

**Khoa: CNTT**

**Bộ môn: TTNT**

**BÁO CÁO TIẾN ĐỘ**

**THỰC HIỆN NGHIÊN CỨU KHOA HỌC**

**GV hướng dẫn:** TS.Ngô Xuân Bách

**Họ và tên:** Lê Thị Ngọc Châm **Lớp:** D13CN6

**Tên đề tài:** ***Phân loại văn bản Tiếng Việt bằng phương pháp NAÏVE BAYES***

1. **Nội dung công việc:**
2. Hoàn thành dữ liệu (chuẩn nhãn, chuẩn title, chuẩn content)
3. Tiền xử lý dữ liệu (bỏ stopword, cut off, v.v)
4. Tách câu, tách từ, gán nhãn từ loại, sử dụng vnsentDetector và vnTagger.
5. Cài đặt phương pháp phân loại NAÏVE BAYES
6. Làm thực nghiệm 10-fold cross-validation, đo kết quả thực nghiệm sử dụng các độ đo Accuracy, Precision, Recall, và F1, tính kết quả trung bình trên 10 folds
7. **Tiến độ thực hiện:**
8. ***Công việc đã hoàn thành:***

* Hoàn thành dữ liệu.
* Tìm hiểu về phương pháp phân loại NAÏVE BAYES.
* Thống kê các từ stopword.
* Tách câu sử dụng vnsentDetector.

1. ***Công việc chưa hoàn thành.***

* Chưa tách từ và gán nhãn từ loại bằng vnTagger.
* Chưa tiền xử lý dữ liệu.
* Chưa cài đặt được phương pháp phân loại Bayes
* Chưa làm thực nghiệm.

1. ***Công việc dự kiến vào tuần tới***.

* Tách từ và gán nhãn từ loại.
* Hoàn thành tiền xử lý dữ liệu
* Tính xác suất của từ.

*Hà nội, ngày 17 tháng 7 năm 2016*

**Sinh viên**

Châm