Project Giữa kỳ

Môn: Nhập môn Xử lý ngôn ngữ tự nhiên

1. **Mục tiêu**

Xây dựng chương trình (tool) gán nhãn từ loại (POS tagger) cho tiếng Việt.

Kết quả đạt được là chương trình, có 2 chức năng:

* Đánh giá độ chính xác trên tập Test: tỉ lệ số từ gán nhãn đúng trên tổng số từ
* Input vào một câu tiếng Việt, cho ra nhãn của từng từ trong câu, ví dụ.

(lưu ý, sử dụng thêm Library để Word Segmentation cho tiếng Việt. Tự tìm Library)

* + Câu input: “Dù khá đắt nhưng tôi vẫn đồng ý.”
  + Kết quả: “Dù/C khá/R đắt/A nhưng/C tôi/P vẫn/R đồng\_ý/V ./.”

1. **Dữ liệu cho trước**

* File traing: “vi\_train.pos”
* File test: “vi\_test.pos”
* File hướng dân thông tin về nhãn POS: “Thon tin nhan tu loai-v2.pdf”

1. **Đánh giá**

**(tổng 11 điểm)**

* (6 điểm) Xây dựng được mô hình thực hiện được mục tiêu, có thể tự code hoặc dùng thư viện hoặc open source, sử dụng Hidden Markov Model.
  + Tự code (1.5 điểm)
* (1.5 điểm )Sử dụng thêm phương pháp phân loại:
  + Phân loại: sử dụng Maximum Entropy Model (Logistic Regression), SVM.
* (2.0 điểm )Trình bày lý thuyết về Conditional Random Fields và sử dụng thư viện (hoặc open source) để giải quyết bài toán trên.

1. **Nộp bài**

Cho các kết quả vào thư mục, nén lại và nộp. Tên thư mục là Họ tên và MSSV

* Chương trình: file họ\_tên\_mssv\_model.jpynb
  + Nếu có nhiều chương trình thì bổ sung thông tim Model vào tên file
* Slide:
  + các kết quả test của các model
  + lý thuyết về CRFs