Nội dung mà em trình bày gồm n phần. Đầu tiên, em xin được giới thiệu vấn đề.

- Là một bài toán có nhiều ứng dụng thực tế
  - phân loại và đánh giá trang web
  - lọc thư rác, tin nhắn rác
  - phân loại sách báo
  - ...
- Chưa có nhiều công cụ, thư viện hỗ trợ phân loại văn bản Tiếng Việt

Trong thời đại bùng nổ thông tin hiện nay, những hướng nghiên cứu về nhận dạng và xử lí văn bản là rất cần thiết, là bước tiền xử lí cho nhiều quá trình để trích xuất thông tin có ích, đặc biệt là các văn bản Tiếng Việt. Tuy nhiên, Tiếng Việt lại chưa nhận được sự hỗ trợ nhiều như Tiếng Anh. Do vậy, nhóm em quyết định thực hiện đề tài này.

Đặc điểm của đề tài

- Vấn đề tách từ cho Tiếng Việt
  - Xâu đầu vào: "Đây là một ví dụ đơn giản minh họa cho việc sử dụng công cụ Đông Du để tách từ."
  - Xâu đầu ra: "Đây là một ví_dụ đơn_giản minh_họa cho việc sử_dụng công_cụ Đông_Du để tách_từ ."
- Vấn đề trích ra tập các từ khóa
  - Xâu đầu vào: "Việt_Nam tái khẳng_định chủ_quyền đối_với hai quần_đảo Hoàng_Sa và Trường_Sa, sau khi Chủ_tịch Trung_Quốc nói rằng các quần_đảo này là "của Trung_Quốc"."
  - Các từ ý nghĩa thấp: tái, đối_với, và, sau, khi, là, . . .
  - Xâu đầu ra: "Việt_Nam chủ_quyền Hoàng_Sa Trường_Sa, Chủ_tịch Trung_Quốc Trung_Quốc"

Trong Tiếng Việt, dấu cách không được sử dụng như một kí hiệu phân tách từ, nó chỉ có tác dụng phân tách các tiếng với nhau. Vấn đề từ đơn, từ ghép.

Tách từ Tiếng Việt

- phương pháp ghép cực đại: sử dụng từ điển, đặt các từ vào câu sao cho phủ hết câu đó
- các phương pháp học máy như đồ thị hóa, hidden Markov, conditional random field, maximum entropy
  - phương pháp Support Vector Machine (SVM) đi kèm pointwise => thư viện Đông Du (C/C++)
  - So sánh dongdu và vnTokenizer

Tiêu chí	vnTokenizer	dongdu
Độ chính xác	97,2%	98,2%
Thời gian (giây)	194,672	26,2
RAM (MB)	19,8	15,1

Bảng: So sánh dongdu và vnTokenizer

ưu điểm: nhanh; nhược điểm: độ chính xác thấp, không xử lí được từ có trong từ điểnSử dụng bản sửa đổi của thư viện Đông Du để tiến hành tách từ trên Python