Trường Đại học Bách Khoa Hà Nội
Viện Công nghệ Thông tin và Truyền thông

BÁO CÁO ĐỒ ÁN MÔN HỌC

Đề tài: PHÂN LOẠI BÀI BÁO TIẾNG VIỆT

Sinh viên thực hiện: Hoàng Văn Khoa
Vũ Mạnh Kiểm, 20111731
Hoàng Văn Khoa, 20159504
Lê Anh Thư, 20112250
Giảng viên: TS. Thân Quang Khoát

Ngày 13 tháng 11 năm 2015

Nội dung

Đặt 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

Phát biểu bài toán

Yêu cầu bài toán

Cho một tập hợp các bài báo, phân loại các bài báo này vào các nhóm đã biết cho trước.

Ý tưởng chung

Sử dụng thuật toán phân loại Naïve Bayes

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"

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

Trích xuất từ khóa

- Xếp hạng từ có ý nghĩa cao hay thấp dựa vào trọng số tf.idf
- Từ có ý nghĩa cao là:
  - Xuất hiện nhiều lần trong một (lớp) văn bản (đồng biến với số lần sử dụng trong (lớp) văn bản)
  - Xuất hiện ít trong các (lớp) văn bản khác (nghịch biến với số (lớp) văn bản sử dụng nó)
- Công thức tính: $w_{tf.idf}(t,d) = w_{tf}(t,d) * idf(t)$
  - $$w_{t,d} = \begin{cases} 1 + \log tf_{t,d} & \text{nếu } tf_{t,d} > 0 \\ 0 & \text{nếu ngược lại} \end{cases}$$
  - $tf_{t,d}$ là số lần từ t được sử dụng trong văn bản d
  - $idf(t) = \log N/df_t$ với N là số văn bản trong bộ dữ liệu, df_t là số văn bản chứa từ t

Tập học và tập kiểm thử

- Kích thước tập học
- Độ phong phú của tập học
- Số từ khóa trích xuất ra
- Đánh giá độ chính xác bằng phương pháp hold-out

Cài đặt

- Ngôn ngữ lập trình chính: Python
- Chỉnh sửa thư viện Đông Du để gọi được hàm tách từ

Kết quả

Thể loại	Số lượng	Độ chính xác
Kinh doanh	98	90,816326531
Sức khỏe - Gia đình	12	91,666666667
Thể thao	77	87,012987013
Thời sự - Chính trị	60	23,333333333
Xã hội	47	59,574468085
Trung bình		70,480756326

Bảng: Kết quả phân loại

Đánh giá và định hướng

Đánh giá

- Đã đạt được
  - Xây dựng được công cụ để phân loại bài báo Tiếng Việt
  - Có thể sử dụng như một thư viện cũng như tương tác qua interpreter của Python
  - Hoạt động được trên Windows, Linux
- Hạn chế
  - Kích thước tập học và tập kiểm thử còn hạn chế
  - Chưa tối ưu được kết quả học
  - Chưa cung cấp dưới dạng giao diện đồ họa GUI

Đánh giá và định hướng

Định hướng

- Thu thập thêm để tăng tập học về kích thước và phong phú về nội dung
- Có thêm bước tối ưu hệ thống
- Cung cấp giao diện đồ họa thân thiện hơn

Tài liệu tham khảo chính

TS. Thân Quan Khoát

Bài giảng Học máy

Viện CNTT-TT, Trường Đại học Bách Khoa Hà Nội, 2015.

TS. Nguyễn Bá Ngọc

Bài giảng Tìm kiếm và trình diễn thông tin

Viện CNTT-TT, Trường Đại học Bách Khoa Hà Nội, 2015.

Lưu Tuấn Anh, Yamamoto Kazuhide

Ứng dụng phương pháp Pointwise vào bài toán tách từ cho Tiếng Việt

NLP Lab, Deparment of Electrical Engineering, Nagaoka University of Technology, 2007.

EM XIN CẢM ƠN
CÁC THẦY VÀ CÁC BẠN ĐÃ CHÚ Ý LẮNG NGHE