BÀI TẬP LỚN
Môn Xử lý ngôn ngữ tự nhiên

Giáo viên: Lê Thanh Hương
Email: huonglt@soict.hust.edu.vn

Dưới đây là các đề bài tập lớn gợi ý cho sinh viên. Sinh viên có thể tự đề xuất đề tài khác nhưng đề tài chỉ được chấp nhận nếu có sự đồng ý của giáo viên.

1. Tìm hiểu cấu trúc hệ thống tìm kiếm thông tin Google hiện tại và các kỹ thuật xử lý trong tìm kiếm thông tin của Google

2. Khai phá dữ liệu văn bản: quyết định một trang web có phải là trang web cá nhân (home page) hay không.

3. Cải tiến phương pháp xác định biên giới câu.

4. Phân tích cú pháp

5. Phân tích ngữ nghĩa: giải quyết vấn đề đồng tham chiếu trong các câu đã được PTCP người

6. Xây dựng chương trình cho phép chuyển đổi các tài liệu dạng văn bản về một lĩnh vực nhất định sang CSDL với các trường dữ liệu đã được xác định sẵn (bởi người thiết kế CSDL). CSDL có thể bằng tiếng Việt hoặc tiếng Anh. Hãy tận dụng các công cụ có sẵn như Gate hay Lucence.
Ví dụ:
a. Thu thập các thông tin liên hệ của các tổ chức có thông tin trên mạng và lưu vào 1 file XML hoặc 1 CSDL gồm có: tên , địa chỉ, số điện thoại, số fax, email. Tiêu chí tìm tổ chức được nhập từ bàn phím, ví dụ, tìm các trường đại học và cao đẳng ở VN, hoặc tìm các công ty tin học ở Hà Nội.
b. Thu thập thông tin về các cửa hàng bán điện thoại di động có thông tin trên mạng và lưu vào 1 file XML hoặc 1 CSDL gồm có: tên điện thoại, hãng, tính năng, giá tiền, nơi bán, địa chỉ, điện thoại liên hệ, email liên hệ.
c. Thu thập thông tin về các hội thảo công nghệ thông tin và lưu vào 1 file XML hoặc 1 CSDL gồm có: tên hội thảo, phạm vi hội thảo (trong nước, quốc tế, châu á,…), địa điểm, thời gian diễn ra hội thảo, địa chỉ trang Web, deadline abstract, deadline fullpaper, acceptation time. Tiêu chí tìm hội thảo được nhập từ bàn phím dưới dạng các từ khoá, ví dụ, call for papers, 2007, 2008, natural language processing.
d. Trích rút tên riêng từ các bài báo tiếng Việt
e. Nhận dạng tên thực thể

7. Tóm tắt đơn văn bản, tóm tắt đa văn bản

8. Phân cụm văn bản

9. Phân loại văn bản:
- phân loại thư, lọc thư rác
- phân loại trang web

10. Dịch máy thống kê hướng miền ứng dụng cụ thể.

11. Tìm kiếm thông tin:

- Đề xuất một số phương pháp cải tiến công cụ tìm kiếm kiểu so khớp và cài đặt

Yêu cầu:

Mỗi nhóm có tối đa 4 người. Đề nghiên cứu lý thuyết từ 1-2 người. Các đề tài có cài đặt chương trình (có thể tận dụng các mã nguồn mở có sẵn) từ 1-4 người. Tất cả các nhóm đều phải báo cáo và demo chương trình (nếu có). Mọi người trong nhóm đều phải tham gia báo cáo phần kết quả của mình.

Về báo cáo:

- Báo cáo cần ≥ 8 trang
- Đối với đề liên quan đến cài đặt chương trình, báo cáo viết dưới dạng tài liệu kỹ thuật có phân tích đánh giá một số hướng tiếp cận liên quan, phân tích phần cài đặt chương trình (các cấu trúc dữ liệu, thuật toán), một số kết quả đạt được, đánh giá độ chính xác và định hướng phát triển.
- Tất cả các báo cáo đều phải chỉ rõ đóng góp của từng thành viên trong nhóm thực hiện đề tài. Báo cáo cần có phần tài liệu tham khảo.

Một số mã nguồn mở về Xử lý ngôn ngữ tự nhiên:

Stanford's Core NLP Suite (viết bằng Java): http://stanfordnlp.github.io/CoreNLP/
Natural Language Toolkit (viết bằng Python): http://www.nltk.org/
Apache Lucene and Solr: http://lucene.apache.org/
Apache OpenNLP (viết bằng Java): http://opennlp.apache.org/
Apache UIMA: https://uima.apache.org/
GATE (General architecture for text engineering, viết bằng Java): https://gate.ac.uk/

Các tài liệu tham khảo:

1. Grant Ingersoll, Thomas Morton, Drew Farris. *Taming Text* : cho người lập trình bắt đầu học NLP và Search. Mỗi chương đều có ví dụ sử dụng các mã nguồn mở.
2. Steven Bird, Ewan Klein, and Edward Loper .*Natural Language Processing with Python* : hướng dẫn sử dụng NLTK qua các công việc như phân loại văn bản, trích rút thông tin, …