Trường đại học bách khoa Hà Nội

Viện công nghệ thông tin và truyền thông

****

Báo cáo Project 2

Đề tài: Xây dựng công cụ tìm kiếm dựa trên Lucene đối với ảnh và văn bản

Giáo viên hướng dẫn: TS. Trịnh Anh Phúc

Sinh viên thực hiện:

Họ và tên: Vũ Mạnh Kiểm

MSSV: 20111731

Lớp: CNTT-TT 2.2

Khóa K56.

Hà Nôi, 05-2014

Contents

I. Nội dung thực hiện

- Nghiên cứu thư viện mã nguồn mở Lucene và xây dựng công cụ tìm kiếm đối với văn bản:
  - Dựa trên nền tảng VNAlalyzer của anh Cao Mạnh Đạt – cựu sinh viên đại học Bách Khoa hà Nội.
  - Xây dựng ứng dụng demo.
- Nghiên cứu thư viện mã nguồn mở Lucene Image và xây dựng công cụ tìm kiếm đối với ảnh:
  - Dựa trên Lucene Image và xây dựng ứng dụng demo.

II. Giới thiệu về thư viện mã nguồn mới Lucene:

1. Lucene là gì?

Lucene là thư viện mã nguồn mở dùng để phân tích, đánh chỉ mục và tìm kiếm thông tin với hiệu suất cao. Lucene được phát triển đầu tiên bởi Doug Cutting được giới thiệu đầu tiên vào tháng 8 năm 2000. Sau đó vào tháng 9 năm 2001 Lucene gia nhập vào Apache và hiện tại được Apache phát triển và quản lý. Lucene được phát triển trên nền tảng hướng đối tượng với ngôn ngữ chính là Java. Hiện tại đã có một số phiên bản trên các nền tảng ngôn ngữ lập trình khác như: .Net, C++, perl …

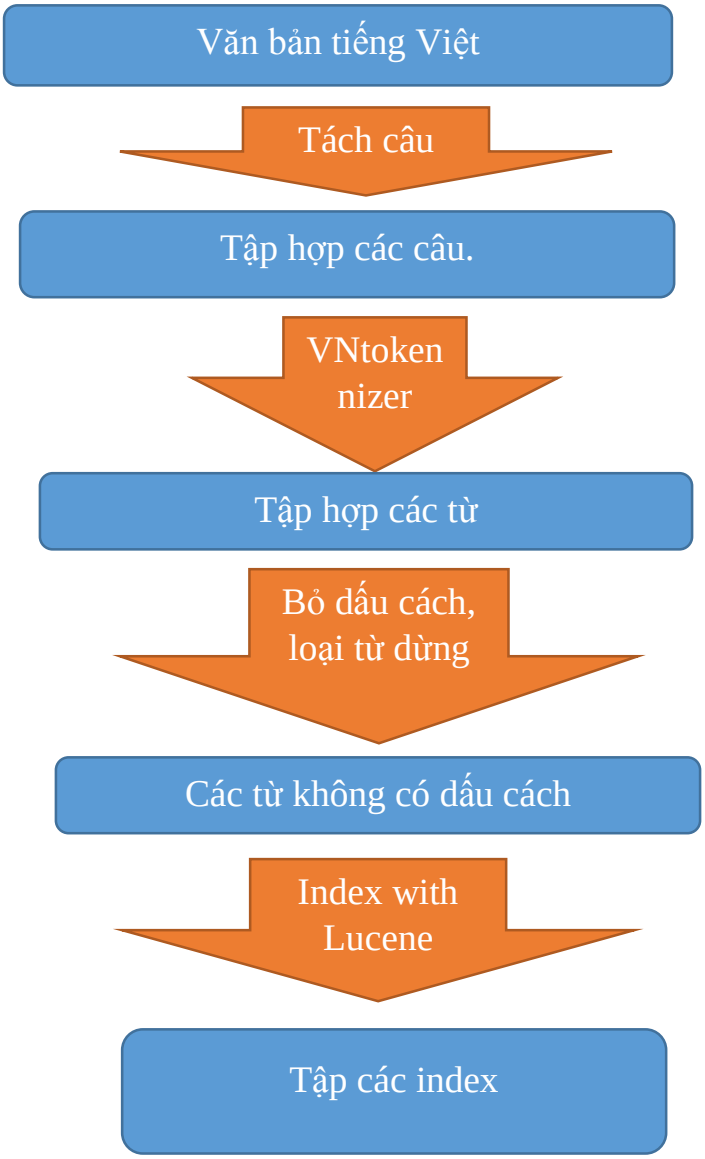
2. Cải tiến Lucene cho phù hợp với ngôn ngữ tiếng Việt.

Lucene hỗ trợ mạnh nhất đối với các thông tin bằng tiếng anh, một số ngôn ngữ khác cũng đã được phát triển trong Lucene nhưng tiếc rằng Lucene chưa hỗ trợ tốt được cho tiếng Việt. Do đó nếu sử dụng trực tiếp Lucene cho các ứng dùng tìm kiếm với nội dung tiếng Việt thì hiệu quả sẽ không cao.

Ý tưởng cải tiến cách sử dụng Lucene cho phù hợp với ngôn ngữ tiêng Việt đã anh Cao Mạnh Đạt đưa ra như sau:

- Sự khác nhau của tiếng anh và Tiếng Việt trong Lucene nằm ở chỗ 2 ngôn ngữ này có cách tách từ khác nhau. Do đó một đoạn văn tiếng Việt được đưa qua Lucene sẽ bị tách thành các tiếng chứ không phải các từ.
- Phương án đưa ra là sử dụng một module tách từ tiếng Việt trước, sau đó thay dấu cách trong các từ nếu có bằng dấu "_" và đưa vào Lucene. Với cách này, Lucene sẽ đánh index đúng cho các từ tiếng Việt như đối với tiếng anh.
- Việc tách từ tiếng Việt được thực hiện thông qua sự trợ giúp của thư viện VNtokenizer của thầy Lê Hồng Phương.

Trong quá trình khi thực hiện đề tài, em đã liên lạc với anh Cao Mạnh Đạt và trao đổi lại một số nội dung trong Project, anh Đạt cũng không còn giữ hoàn chỉnh Project mà chỉ còn các nội dung chính và thư viện đã được build trên Github nên em đã xây dựng lại phần xử lý stopword của project đó.



Hình 1: Quá trình tạo index với văn bản tiếng Việt

3. **Xây dựng công cụ tìm kiếm văn bản tiếng Việt dựa trên Lucene.**
   Em đã sử dụng ý tưởng và một phần code trong project của anh Cao Mạnh Đạt để xây dựng tiếp nên công cụ tìm kiếm của mình. Công cụ này sử dụng tìm kiếm nội dung trong văn bản. Có 2 điều mới mà em đã bổ sung so với Project của anh Cao Mạnh Đạt, đó là:

- Khả năng tìm kiếm trong file word và file PDF. Project ban đầu chỉ hỗ trợ tìm kiến trên file text (.txt), em đã bổ sung mở rộng tìm kiếm trên file .doc và .pdf
- Khả năng tìm kiếm nâng cao khi người dùng có câu truy vấn lỗi chính tả. Việc so khớp trong Lucene là so khớp chính xác do đó nếu người dung nhập từ khóa lỗi thì sẽ không thể tìm ra kết quả. Em đã xây dựng module so sánh đưa ra các kết quả với các từ có độ sai khác không quá 30% so với từ người dùng nhập vào. Do đó người dùng vẫn có khả năng tìm ra văn bản nếu không may nhập sai.

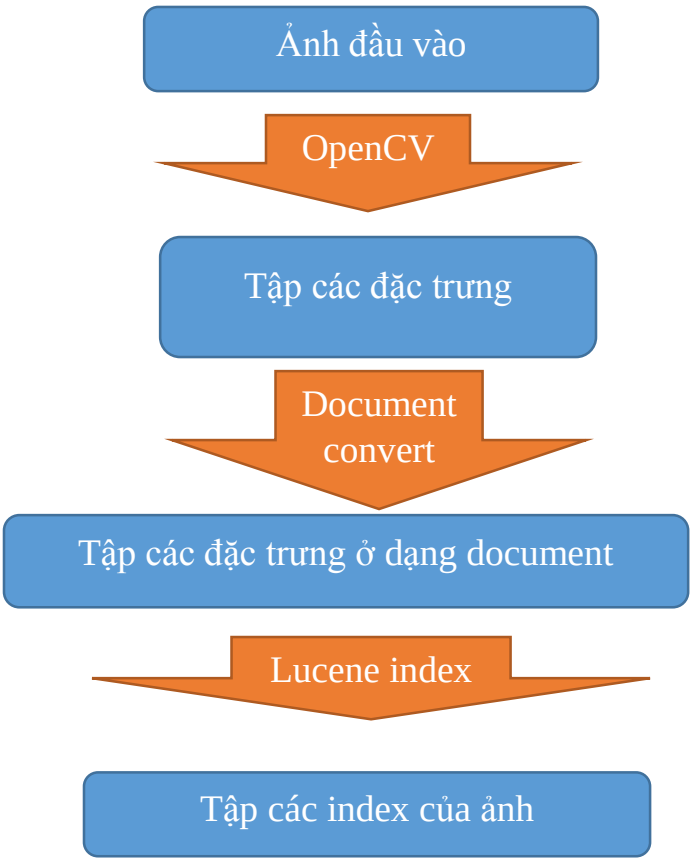
III. Giới thiệu thư viện mã nguồn mở Lucene Image

1. Lucene image là gì?

Lucene Image là thư viện tìm hỗ trợ các API cho việc đánh index và tìm kiếm trên ảnh. Lucene Image được xây dựng dựa trên nền tảng của Lucene để đánh index và kết hợp với thư viện OpenCVđể xử lý, trích xuất dữ liệu từ ảnh.

Quán trình xử lý của thư viện Lucene Image có thể được miêu tả như sau:

- Ảnh đầu vào được đưa qua thư viện OpenCV để trích chọn đặc trưng.
- Các đặc trưng của ảnh đã trích chọn sẽ được chuyển đổi sang dạng document (dữ liệu dạng text).
- Dùng Lucene đánh index cho dữ liệu đã được chuyển đổi sang dạng document trên.



Hình 2: Quá trình đánh index với ảnh trên Lucene Image

2. **Ứng dụng Lucene Image để xây dựng công cụ tìm kiếm ảnh đơn giản.**
   Ứng dụng tìm kiếm ảnh của em dựa trên Lucene Image chỉ sử dụng các đặc trưng toàn cục của ảnh là histogram (sự phân phối màu sắc của ảnh) để đánh index và so sánh giữa các ảnh.

   Đặc điểm của việc dùng các đặc trưng toàn cục đó là giúp cho việc sử lý và đưa ra phản hồi nhanh hơn. Tuy nhiên nó có hạn chế về độ chính xác. Khi ảnh có sự thay đổi về kích thước hay ảnh này là một phần của ảnh khác thì nó không có hiệu quả nhiều. Khi 2 ảnh có nội dung khác nhau nhưng có sự phân bố màu sắc giống nhau thì nó có thể trả về 2 ảnh này tương đồng nhau.

   Kết quả test thực tế với bộ ảnh nhỏ của ứng dụng cũng chỉ cho độ chính xác khoảng 30%. Để nâng được độ chính xác thì sẽ phải sử dụng các đặc trưng cục bộ, và điều này sẽ đòi hỏi phải hy sinh tiêu chí thời gian phản hồi.

Hiện em đang xây dựng một module tìm kiếm ảnh dựa trên đặc trưng cục bộ nhưng nó chưa hoàn thành để có thể sử dụng. Tuy nhiên độ chính xác của nó sẽ tăng lên đáng kể.

Tài liệu tham khảo:

1. Thư viện Lucene: https://lucene.apache.org/core/
2. VNAnalyzer – Cao Mạnh Đạt:
   https://caomanhdat.wordpress.com/2013/06/26/bo-phan-tich-tu-vung-tieng-viet-cho-lucene/
   Link project:
   https://github.com/CaoManhDat/VNAnalyzer
3. Thư viện Lucene Image: http://www.lire-project.net/

Lời cảm ơn

Em đã rât hứng thú khi tìm hiểu và sử dụng thư viện Lucene trong suốt quá trình hoàn thành môn học Project 2 này. Em đã khám phá ra một số điều mới lạ và một số cách suy nghĩ mới khi giải quyết một số bài toán thực tế.

Môn học đã đem lại cho em rất nhiều về kỹ năn tìm hiểu tài liệu, thực hành và tạo ra ứng dụng thực tế. Em rất cảm ơn thấy Trịnh Anh Phúc đã hướng dẫn và giúp em hoàn thành nội dung tìm hiểu.

Em xin chân thành cảm ơn!

Hà Nội, 5-2015

Sinh viên

Vũ Mạnh Kiểm