**LUCENE IN ACTION**

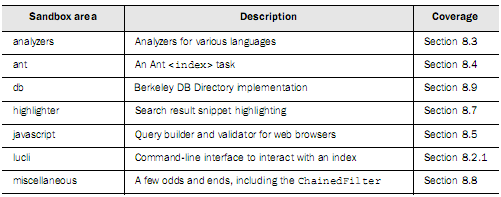
# CHAPTER 8: Tools and extensions

Trong chương này giới thiệu những công cụ hỗ trợ index , browse hay query. Có một vài bộ phân tích (analyzer) đặc trưng cho nhiều ngôn ngữ có trong Sandbox của Lucene. Cân nhắc chú ý dùng những phần source code đi kèm theo phần thiết kế Lucene vì có cuốn sách khác nhau về Lucene có sự riêng biệt về các công cụ hỗ trợ được phát triển ngày càng nhiều và mới vì Lucene là mã nguồn mở.

## Playing in Lucene’s Sandbox

Ngày nay , các đóng góp cung cấp phát triển mã nguồn mở liên quan đến Lucene project đã và đang ngày càng tăng lên một cách nhanh chóng, vì vậy kho chứa Sandbox CVS đã được tạo ra.Sandbox liên tục được phát triển , đưa ra và phát triển những mẫu dự án cụ thể và thực hiện có định hướng rõ ràng. Khi cần những gì cần thiết với những phần liên quan đến Lucene có thể tìm kiếm trong Sandbox , sẽ có rất nhiều kết quả tìm thấy để hỗ trợ , và nó là mã nguồn mở nên bạn cũng có thể ý cung cấp chia sẽ mã nguồn mở liên quan đến Lucene.

Sau đây là những nội dung chính hiện có trong Sandbox tại thời điểm hiện tại viết cuốn Lucene in action:

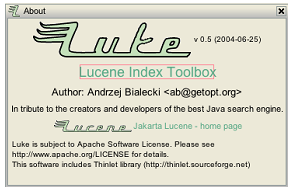


## Interacting with an index

Bạn có thể tự tạo ra một index hay bằng cách viết bằng code Java tương tác với Lucene áp dụng vào chương trình của bạn.Có một vài tiện ích đã được tạo ra cho bạn sử dụng tương tác với Lucene trong sử lý index:

■ lucli (Lucene Command-Line Interface)—là một giao diện command line giúp truy vấn đặc biệt và kiểm tra index. Lucli được đóng góp từ Dror Matalon, nó cung cấp những dòng lệnh thực thi , cho phép xem lại những dòng lệnh đã thực thi, quay lại thời điểm thực thi những dòng lệnh trước đó.Lucli tương tác với một chuỗi hoạt động liên tiếp . Nó có 2 hạn chế đã được ghi chép lại nhưng nhìn chung thì vẫn không làm giảm sự hữ dụng của nó : phiên bản hiện tại nó cho sử dụng MultiFieldQueryParser cho tìm kiếm và hard-coded sử dụng StandardAnalyzer với parser.

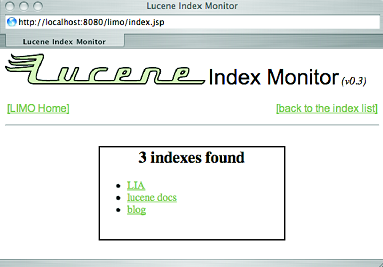
■ Luke (Lucene Index Toolbox)—Một chương trình giao diện desktop với các tiện ích hay . Andrzej Bialecki đã tạo ra Luke (found at <http://www.getopt.org/luke/> ). Luke trực quan đối với người sử dụng cho index browser.



Tiện ích Luke cho phép hỗ trợ đối với những query đặc biệt , cung ứng những terms và structure từ index. Luke dần trở thành công cụ chuẩn cho phát triển Lucene, giúp user có thể sử dụng browsing và thử nghiệm nhanh. Được cài đặt từ gói file java , là Java WebStart nhưng được cài đặt trên localhost.

Có 5 tab trực quan trong Luke : Overview, Documents, Search, File, Plugins.Các chức năng được phân chia rõ ràng tường minh cho người dùng dễ dàng sử dụng. Ví dụ như trong tab Overview khi sử dụng cần điều chỉnh lại đường dẫn file index và những file được chọn để thực hiện index. Index có những điểm hay khi sử dụng Luke như cho phép thống kê “Top ranking terms”, có thể xem những tài liệu chứ 1 term được phân tích và hiển thị , … Khi Search cho hiện kết quả , tần xuất xuất hiện trong từng văn bản tìm được, những term được chứa ,… Ngoài ra còn cho hỗ trợ sử dụng những plug-in cho Luke cho sử dụng những chức năng mới.

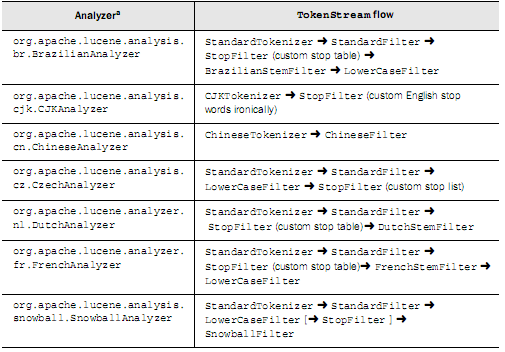
■ LIMO (Lucene Index Monitor)—Giao diện web cho phép remote index browsing. Julien Nioche là tác giả của Lucene Index Monitor(LIMO) , link <http://limo.sourceforge.net/> , là một giao diện web để tương tác Lucene index . Từ link trên có thể tải source code và cài đặt cho localhost , có thể dùng Tomcat để thực thi source code , ngoài ra cần phải thêm thư viện liên quan từ Lucene , sử lại nội dung trong file limo/WEB-INF/web.xml cho phù hợp. Sau khi cài đặt xong sẽ chạy với localhost : <http://localhost:8080/limo/index.jsp>



Phần sử dụng bạn có thể tự tìm hiểu .

## Analyzers, tokenizers, and TokenFilters,oh my

Có nhiều bộ Analyzer để tham khảo , nhưng Sandbox không thể làm thỏa mãn hết các yêu cầu theo từng vùng . Nó hỗ trợ một vài ngôn ngữ thông dụng cho bộ analyzer , bộ lọc , tokenizers và thuật toán trong Snowball của analyzer. Một số bộ phân tích được tóm lược trong bảng sau :



* SnowballAnalyzer

Có những điểm nổi bật như có thể tách ra như một phần driver của một tổng thể để phân chia cho các ngôn ngữ khác nhau. . Nếu trong project của bạn có nhu cầu về stremming , đầu tiên bạn có thể nghĩ ngay đến Snowball analyzer , chắc chắn bạn phải custom lại nó để có thể sử dụng theo nhu cầu của project của bạn.

* Obtaining the Sandbox analyzers

Tùy thuộc vào những cái bạn cần , bạn có thể sử dụng các gói bộ phân tích trong bộ thư viện để mượn những ý tưởng và tạo ra một bộ phân tích mới tùy ý.

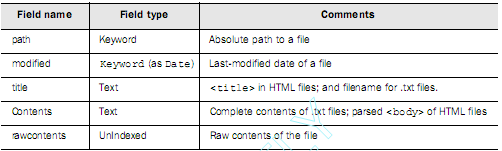
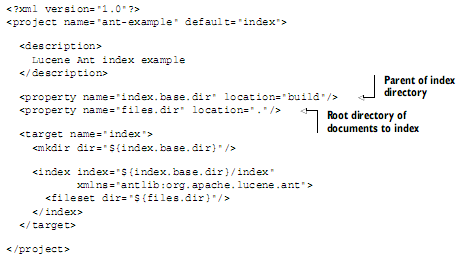
## Java Development with Ant and Lucene

Một điểm tự nhiên với Lucene đó là sự kết hợp các tài liệu indexing trong suốt quá trình build. Vì dữ liệu thường xuyên được thay đổi , vì vậy cần phải tìm kiếm dữ liệu trên những phiên bản mới nhất của dữ liệu được cập nhật. Java Development with Ant( Hatcher and Loughran, Man-

ning Publications, 2002), Erik created an Ant task to index a directory of file-based .Và code này đã được tăng lên và duy trì trong Sandbox.

* Using the <index> task

Ví dụ:



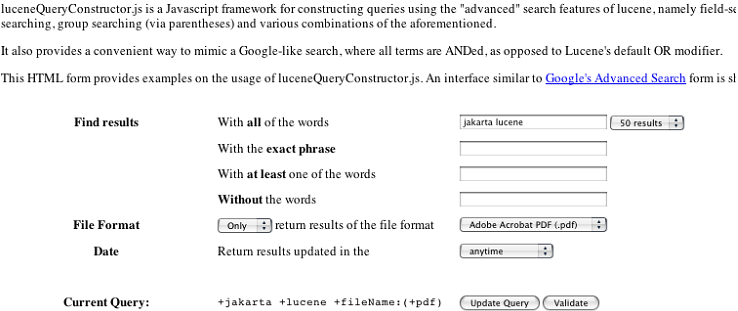
Mặc định những file xử lý làm .txt hoặc .html . Nếu như chỉ sử dụng mặc định xử lý trên những document thì không đủ , nhưng có thể custom document handler .

## JavaScript browser utilities

QueryParser là một tiện ích , nó dễ dàng đưa ra những textbox đơn giản đồng ý cho phép user gõ vào một truy vấn, nhưng nó không thân thiện khi sử dụng đối với user để có thể thấy được tùy chọn truy vấn riêng biệt từ field , như là lựa chọn sắp xếp theo ngày để kết hợp với text box để tìm kiếm. Tiện ích JavaScript trong Sandbox hỗ trợ với tiện ích browser-side trong việc xây dựng và làm tương hợp với QueryParser.

* JavaScript query construction and validation

Trong những chương trước có thể nhận thấy rằng QueryPaser đưa ra trực tiếp cho end user sử dụng và có thể đưa đến sự hỗn độn. Nếu bạn cung cấp một trang web tìm kiếm , bạn có thể nghĩ đến việc sử dụng JavaScript query trong Sandbox . javascript Sandbox bao gồm các file ví dụ HTML giống như tìm kiếm nâng cao của Google với các lựa chọn khi search cho user sử dụng.Hỗ trợ tất cả các field HTML gồm text, hidden text , radio buttons, single và multiple selects. Với từng field HTML cần phải có tương ứng với tên đó được gắn thêm hậu tố *Modifier* .Modifier field có thể là một hidden field để ngăn cản user có thể điều khiển được nó . JavaScript file cho tùy biến với những đặc điểm giống như debug thông báo cho bạn biết những cái gì đang xảy ra. Rất dễ dàng để sử dụng JavaScript . JavaScript Sandbox đã được tăng lên rất nhiều. Có nhiều tiềm năng trong HTML , chúng ta có thể xem các ví dụ trong Sandbox khi cần biết sự hữu dụng của nó.



* Escaping special characters

QueryParser sử dụng nhiều ký tự đặc biệt , vì vậy những ký tự đó cần phải được tránh nếu như chúng sử dụng trong field name hoặc một phần của term. Sử dụng lucene- QueryEscaper.js hỗ trợ từ Sandbox , bạn có thể escape một query string.Bạn chỉ nên sử dụng query escape đối với fields và strings mà nó không chứa bất kỳ một ký tự đặc biệt nào.

* Using JavaScript support

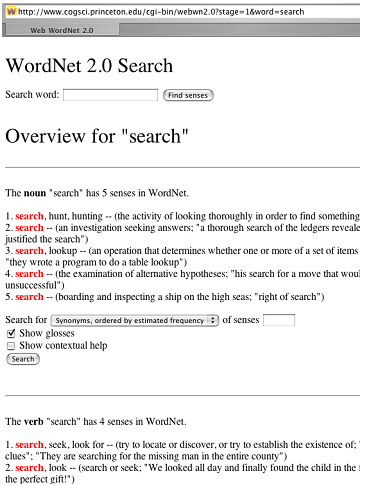
Thêm JavaScript hỗ trợ file HTML có thể thêm đường dẫn file js vào trong thẻ <head> như



Gọi hàm doMakeQuery để xây dựng một truy vấn và doCheckLuceneQuery để xác nhận câu truy vấn. Phương thức require một đối số của form field được định rõ nơi nào được đưa ra và thông qua. Escape query string sẽ bị gọi trở lại.

## Synonyms from WordNet

Có một hệ thống phát triển tại Princeton Univer-sity’s Cognitive Science Laboratory, driven by Psychology Professor George Miller minh họa hệ thống từ đồng nghĩa. WordNet được hình dung là dạng của từ ngữ có thể chuyển đổi cho nhau vả về từ vựng và ngữ nghĩa. Theo định nghĩa của Google thì đặc trưng thường nói đến những user trực tuyến hệ thống WordNet, đồng ý cho thông qua từ nối liền với nhau. Có thể xem hình ảnh sau cho hiện kết qyar tìm kiếm từ “search” tại WordNet.



Dữ liệu về các từ đồng nghĩa trong WordNet có thể được sử dụng trong Lucene Index. Nó đồng ý cho tìm kiếm nhanh những từ đồng nghĩa được đưa vào trong suốt quá trình indexing hoặc querying.

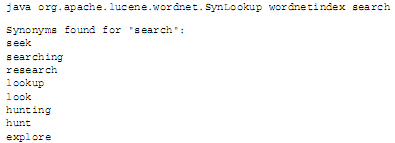
* Building the synonym index

Để cài đặt cho synonym index cần download và giải nén source code từ WordNet tại trang <http://www.cogsci.princeton.edu/~wn> , sau đó xây dựng synomym index sử dụng Syns2Index từ command line.Gồm 2 đối số : đối số 1 là đường dẫn đến file wn\_s.pl và đối số 2 là đường dẫn nơi Lucene index sẽ được tạo.



Sử dụng cơ sở dữ liệu từ WordNet để Lucene index field *word* và không index lại field đối với những từ đồng nghĩa (*syn*) cho từng dữ liệu đưa vào. Phiên bản 1.6 của WordNet có thể làm việc 39.718 documents và thực hiện với từng từ trong từng dữ liệu và cho ra file index với kích cỡ 2.5MB , với kích thước này nó có thể đủ để nạp vào RAMDirectory truy cập với tốc độ nhanh chóng.

Một tiện ích thứ 2 của chương trình WordNet Sandbox là giúp tìm kiếm những từ đồng nghĩa của một từ cho trước.



* Tying WordNet synonyms into an analyzer

Trong phần custom SynonymAnalyzer có thể dễ dàng kết hợp với WordNet sử dụng SynonymEngine . Có thể sử dụng với SynonyAnalyzer, điều chỉnh SynonymAnalyzerViewer sử dụng WordNetSynonymEngine .

Lucene WordNet có source code trên Lucene Sandbox tại địa chỉ <http://jwn.sourceforge.net/>

* Calling on Lucene

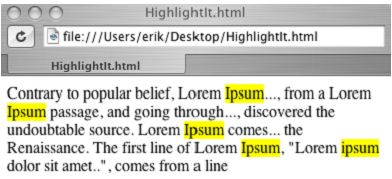
Lucene cpnf được dùng trong điện thoại , do yêu cầu trong điện thoại có kích thước bộ nhớ nhỏ nên kích thước của WordNet cũng được thu nhỏ lại cho phù hợp .T9 interface hiện tại được phổ dụng nhất trong phone.

Việc tìm kiếm trên màn hình điện thoại di động thì nhanh và đúng đắn thuyết phục . Index trong Lucene thì sử dụng hơn 2MB không phù hợp lắm với bộ nhớ của một số mobile-phone hiện tại .

## Highlighting query terms

Cho người sử dụng có thể tìm kiếm một vài nội dung xung quanh những thứ cần tìm và có liên quan với nó, thực sự điều đó thì rất hữu dụng.Với từng kết quả , có thêm những highlighting đối với những terms của truy vấn , nó sẽ tóm tắt được những nội dung bao quanh tìm kiếm những term truy câu truy vấn , và như thế cũng đủ để biết được và đánh giá sự đúng đắn của kết quả tìm kiếm .

Mark Harwood đã đóng góp vào Sandbox về highlight text . Ví dụ về highlighting với term truy vấn là *ipsum*.



Code về Highlight gần đây đã có những bước phát triển vững chắc trở thành một tiện ích tinh vi và linh động. Hightlighting gồm 3 phần chính: Fragmenter, Scorer, Formatter. Analyzer tạo ra TokenStreams, term trong Query thì được kết nối với Tokens đưa ra từ TokenStream. Những đoạn text nào giống với term ở trong đoạn text gốc thì được highlight .

Đi kèm với SimpleFragmenter chia cắt đoạn text gốc thành những đoạn fragments có kích thước giống nhau với mặc định là 100 ký tự. Highlight sẽ thực hiện lần lượt với từng đoạn.

QueryScore đi kèm trong Scorer .Công việc chính của Scorer sắp xếp thứ hạng của fragments. QueryScorer sử dụng terms từ truy vấn (Query) .

Formatter trang trí cho term text , làm nổi bật với những đoạn text trùng với dữ liệu được đưa vào truy vấn.

* Highlighting with CSS

Sử dụng tag <B> xung quanh đoạn text được highlight trước khi đưa ra cho người dùng xem hoặc có thể dùng CSS để thay thế . Ví dụ custom thẻ bắt đầu và kết thúc để highlight term dùng một thẻ <span> sử dụng CSS với class highlight. Sử dụng các thuộc tính của CSS , color , forrmating cho phần được highlight .

* Highlighting Hits

Những nơi chứa field text ban đầu trong index thì để cho bạn quyết định . Nếu đoạn text ban đầu không chứa trong index nó sẽ cho bạn quyết định lấy lại đoạn text được highlight từ gốc ban đầu. Nếu text được chứa trong field, nó có thể được lấy lại trực tiếp từ Document lấy từ Hits.

## Chaining filters

Sử dụng bộ lọc cho dữ liệu đầu vào để làm hẹp không gian tìm kiếm từ một truy vấn. Sandbox có chứa một bộ lọc đáng chú ý meta-filter được đóng góp bởi Kelvin Tan , nó có thể chuyển móc nối các bộ lọc lại với nhau và thi hành với AND , OR, ANDNOT có thể phối hợp với nhau. Dựa và những mục tiêu cần làm mà bạn có thể tạo ra những filter theo ý muốn để sử dụng.

## Sorting an index in Berkeley DB

Dự án không quá sôi nổi của Chandler (<http://www.osafoundation.org> ) được đưa ra để cố gắng xây dựng một mã nguồn mở Personal Information Manager. Chandle nhắm đến việc quản lý các loại thông tin như email, messages, appointments, contacts, tasks, notes, web pages, blogs, bookmarks, photo, .v.v. Điều bí mật phía bên trong của Chandler đó là sử dung Sleepycat’s Berkeley DB trong một không gian lớn có thể kết nối cơ sở dữ liệu thông qua nhiều con đường khác nhau. Chandler có code sử dụng ngôn ngữ Python.