Vấn đề trong tìm hiểu về Fuzzy Search

Họ và tên: Nguyễn Đức Toàn

Thư viện tìm hiểu : Apache – Lunce

Link tham khảo: <http://www.ibm.com/developerworks/library/os-apache-lucenesearch/>

Giới thiệu :

Lucene là một mã nguồn mở, là 1 thư viện mở rộng khả năng tìm kiếm có sẵn của Apache Software Foundation . Bạn có thể sử dụng Lucene trong ứng dụng mã nguồn thương mại và cởi mở. API mạnh mẽ Lucene tập trung chủ yếu vào chỉ mục văn bản và tìm kiếm . Nó có thể được sử dụng để xây dựng khả năng tìm kiếm cho các ứng dụng như e- mail, danh sách gửi thư , tìm kiếm web , tìm kiếm cơ sở dữ liệu.

Xây dựng ứng dụng với Lunce: xem link <http://www.ibm.com/developerworks/library/os-apache-lucenesearch/>

Demo.

Theo như tìm hiêu được, ta phải tạo 1 index, mình sẽ demo trực tiếp trên từ điển của nhóm:

* Đầu tiên mình phải tạo indexdata . Ở đây mình tạo indexdata nằm trong thư mục sdcard để khi chỉ tạo 1 lần mình tái sử dụng lại. Có thể tạo indexdata trong RAM nhưng với ứng dụng từ điển của nhóm mỗi lần tạo indexdata là rất lâu nên không thể lưu trong RAM.
  + Tạo Indexwirter:

File f = **new** File(Environment.*getExternalStorageDirectory*().getPath() +"/myindex/"); // tạo index nằm trong sd card với directory la myindex

f.mkdir();

Directory fsDirectory = FSDirectory.*getDirectory*(f);

Analyzer standardAnalyzer = **new** StandardAnalyzer();

// Create a new index

**boolean** create = **true**;

// Create the instance of deletion policy

IndexDeletionPolicy deletionPolicy = **new** KeepOnlyLastCommitDeletionPolicy();

indexWriter = **new** IndexWriter(fsDirectory, standardAnalyzer,

create, deletionPolicy,

* + IndexWriter.MaxFieldLength.*UNLIMITED*);

Thực hiện indexdata:

InputStream is = context.getResources().openRawResource(

R.raw.*engvietdictindex*);

BufferedReader reader = **new** BufferedReader(**new** InputStreamReader(is));

String str = "";

**boolean** flag\_dangdoc = **true**;

**while** (flag\_dangdoc) {

str = reader.readLine();

**if** (str == **null**) {

flag\_dangdoc = **false**;

**break**;

}

String[] list = str.split("\t");

String strword = list[0];

String strindex = list[1];

String strlen = list[2];

Field wordfield = **new** Field("word", strword, Field.Store.*YES*,

Field.Index.*ANALYZED*);

**if**(strword.toLowerCase().indexOf("pune") != -1){

// Display search results that contain pune in their subject first by setting boost factor

wordfield.setBoost(2.2F);

}

Field indexfield = **new** Field("idnex", strindex, Field.Store.*YES*,

Field.Index.*ANALYZED*);

Field lenfield = **new** Field("len", strlen, Field.Store.*YES*,

Field.Index.*ANALYZED*);

Document doc = **new** Document();

doc.add(wordfield);

doc.add(indexfield);

doc.add(lenfield);

indexWriter.addDocument(doc);

}

indexWriter.optimize();

indexWriter.close();

Ở bước indexdata trong demo chạy rất ok.

Nhưng khi sử dụng IndexSearcher để đọc cái index:

IndexReader indexReader = IndexReader.*open*(Environment.*getExternalStorageDirectory*().getPath() + "/myindex");

indexSearcher = **new** IndexSearcher(indexReader);

thì mình luôn bị văng ra tai indexSearcher = new IndexSearcher(indexReader).

Điều đáng nói là với code này mình ánh xạ qua java chạy rất ok.

Mình đưa vấn đề này lên để mọi người cùng giải quyết giúp, mình resolve nó mấy ngày nay rồi.

Lưu ý: Ở demo android mình đưa cho các bạn nếu chạy lần đầu bạn Crtl + Shift + / phần code từ number 79 🡪 90 và bỏ phần comment từ line number 156 🡪167. Sau đó chạy ứng dụng và nhấn Create Fuzzy.

Còn nếu chạy lần thứ 2 thì bỏ phần comment từ line number 79 🡪90 và Crtl + Shift + / phần code từ number 156 🡪 167. Sau đó chạy ứng dụng.