

Wittawat's Blog

The place to share what I know

[Home](#)[My Calendar](#)[Photos](#)[About me](#)

การเอาค่าที่ตัดได้จาก Lucene Analyzer

Sep 26 2009 Published by [admin](#) under [Uncategorized](#)

มีผู้อ่านท่านหนึ่งถามมาว่าจะเอาค่าที่ตัดได้จาก Analyzer ของ Lucene ไปใช้ได้อย่างไรจากการลองซึกพิกหนึ่งทำให้ได้วิธีมาดังนี้ครับ

```
import java.io.IOException;
import java.io.StringReader;
import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.Token;
import org.apache.lucene.analysis.TokenStream;
import org.apache.lucene.analysis.th.ThaiAnalyzer;

/**
 * Demonstrate how to obtain a list of Lucene tokens.
 * @author Wittawat Jitkritum (Sep 26, 2009)
 */
public class TestLuceneAnalyzer {

    public static void tokenize(Analyzer analyzer, String text) throws IOException{
        // เรียก method ชื่อ tokenStream เพื่อให้ได้สายของ token มา argument แรกเป็นชื่อ field ใน
        // document ของ Lucene ในที่นี้เราไม่ได้ใช้ ก็ไม่ต้องสนใจ
        TokenStream ts = analyzer.tokenStream(null, new StringReader(text));
        // สร้าง Token ไว้เก็บผลลัพธ์ของค่าที่ได้มา
        Token token = new Token();
        // วน loop เพื่อเอา token ที่ละอันจาก TokenStream เหตุที่ใส่ token เป็น argument เพื่อให้ด้วย
        // เข้าใจว่าอยากให้อ่านจะได้ไม่ต้องสร้างอันใหม่
        while((token = ts.next(token)) != null){
            String term = token.term(); //เรียก method term() บน token เพื่อเอาค่าเป็น String
            System.out.print(term);
            System.out.print(" | ");
        }
        System.out.println();
    }

    public static void main(String[] args) throws Exception{
        Analyzer thaiAnalyzer = new ThaiAnalyzer();
        String text = "นี่คือประโยคเอาไว้ทดสอบการตัดคำของ Lucene Analyzer";
        tokenize(thaiAnalyzer, text);
    }
}
```

พอรัน main ด้านบนก็จะได้ประมาณนี้ครับ

นี่ คือ | ประโยค | เอา | ไว้ | ทดสอบ | การ | ตัด | คำ | ของ | lucene | analyzer |

สำหรับคนที่อยากรู้เรื่องการทำ index โดยใช้ Lucene ขอเชิญคลิกอ่าน [post](#) อันเก่านี้เขียนจบแล้วเพิ่งนึกได้ว่าผู้ที่ถามมาอาจจะอยากเข้าถึงคำที่อยู่ใน index ีป่าว.. ไม่ใช่เอา Analyzer มาใช้เฉยๆแบบนี้ ถ้ากรณีนั้นเอาไว้มีเวลาจะเขียนทีหลังครับ

Free Search facebook
100% Free Emoticons And Smileys to
Download Now For Your Messenger

สมัครงาน บ. ไทยเนฟเวอร์
เปิดรับหลากหลายตำแหน่งวันนี สมัครฟรีและรับหนังสือ
มือทำงาน

โฆษณาโดย Google

3 responses so far

Thai Text Search with Lucene

Jan 19 2009 Published by [admin](#) under [Uncategorized](#)

หลายวันที่ผ่านมามีเพื่อนนักศึกษาปี 4 ต่างมหาวิทยาลัยเมสมาถามเรื่องการหา search engine และสนใจโปรแกรม [Sansarn Offline](#) ก็เลยถือโอกาสนี้แบ่งปันความรู้เรื่องการหา search อย่างง่ายที่นี้ละกัน

Sansarn Offline เขียนโดยใช้ภาษา Java และตัว engine ที่ทำ search นั้นก็ใช้ [Lucene](#) ซึ่งเป็น library สำหรับใช้ในการค้นคืนเอกสารโดยเฉพาะ ต่อไปนี้เป็น code ตัวอย่างการใช้งาน Lucene กับข้อความภาษาไทย สิ่งที่ต้องใช้มี 2 อย่างคือ

1. Lucene library สามารถหาได้จาก lucene.apache.org เมื่อโหลดมาแล้วจะมีไฟล์ชื่อประมาณ lucene-core-2.x.x.jar ให้เพิ่มไฟล์นี้ลงไป Java classpath ด้วย
2. ThaiAnalyzer เป็นตัวตัดคำภาษาไทยเพื่อให้ Lucene ตัดคำไทยได้ จะได้ค้นคืนโดยใช้คำสำคัญ (keyword) ได้ สามารถหาได้จาก <http://sansarn.com/look/download.html> เมื่อได้มาแล้วให้เพิ่ม ThaiAnalyzer.jar ลงใน classpath

เมื่อได้ 2 อย่างที่กล่าวมาแล้วก็เริ่ม demo กันเลย

```
import java.io.IOException;
import org.apache.lucene.analysis.Analyzer;
```

"I know that I know nothing" -- Socrates

Tags

allnews browser conference database data recovery
dynamic dns experiment firefox font fun gadget
gimp gwt hosting **iphone** **java** javascript
jtcc kicss2008 laptop latex library linux
lucene lyric myself mysql netbeans news nlp
parser plugin python qast quiz semantic web
senior project server Song tarot tcc thai trip
ubuntu work

Archives

- ✿ July 2010
- ✿ June 2010
- ✿ May 2010
- ✿ March 2010
- ✿ February 2010
- ✿ January 2010
- ✿ November 2009
- ✿ October 2009
- ✿ September 2009
- ✿ August 2009
- ✿ July 2009
- ✿ June 2009
- ✿ May 2009
- ✿ April 2009
- ✿ March 2009
- ✿ February 2009
- ✿ January 2009
- ✿ December 2008
- ✿ November 2008
- ✿ October 2008
- ✿ August 2008
- ✿ July 2008
- ✿ June 2008
- ✿ May 2008
- ✿ April 2008

Recent Comments

- ✿ admin on การเรียนรู้ด้วยวิธี linear least squares
- ✿ Jirach on การเรียนรู้ด้วยวิธี linear least squares
- ✿ Chotika on การเรียนรู้ด้วยวิธี linear least squares
- ✿ admin on การเรียนรู้ด้วยวิธี linear least

```

import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.analysis.th.ThaiAnalyzer;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.index.CorruptIndexException;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.queryParser.QueryParser;
import org.apache.lucene.search.Hits;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.search.Query;
import org.apache.lucene.store.Directory;
import org.apache.lucene.store.FSDirectory;

/**
 * Demonstrate how to use basic features of Lucene
 * with THai text.
 *
 * You can freely modify this file to suit your needs.
 * There is no any license nor any guarantee associated
 * with this file.
 *
 * @author Wittawat Jitkrittum
 * January 19, 2009 10:30 am
 */
public class LuceneDemo {

    /**In this example, we add only 2 documents.*/
    public static void addSampleDocuments(IndexWriter writer) throws
    CorruptIndexException, IOException{
        /**In Lucene, to add contents to the index, you need to
        construct a Document object. Most of the time, a document
        is one web page. (This is not necessary. It can be one paragraph
        * if you want)*/
        Document doc1 = new Document();
        /**One document consists of a collection of fields. You can think of
        Field as a part of a document. In most cases, we use three fields namely,
        title field, content field, and url field. You may add more if you want.*/
        doc1.add(new Field(
            "title", //this is the field name
            "Test Lucene.", //the title field
            Field.Store.YES, //store the title so that we can retrieve and show later.
            Field.Index.TOKENIZED //Tokenize and index the title so that we can search by
            keywords.
        ));

        /**Add a content field*/
        doc1.add(new Field(
            "content",
            "ผมสร้างไฟล์ขึ้นมาเพื่อเป็น ตัวอย่างแก่ผู้ที่ต้องการสร้าง search engine อย่างง่าย โดยใช้ Lucene" +
            "Lucene เป็น library ที่เข้าใจง่าย ผู้ใช้ไม่จำเป็นต้องรู้รายละเอียดของการทำงานของ search engine
            มาก",
            Field.Store.COMPRESS, //Compress the content. Most of the time, contents is big.
            Field.Index.TOKENIZED
        ));

        /**Add a url field*/
        doc1.add(new Field(
            "url",
            "http://www.thisisasampleurl.com/doc1.html",
            Field.Store.YES,
            Field.Index.NO //There is no need to index nor tokenize the url. We only keep it
            to show to the users. We don't search the url.
        ));

        //-----
        /**Make another document*/
        Document doc2 = new Document();

        doc2.add(new Field("title", "หัวข้อของเอกสารที่ 2", Field.Store.YES,
            Field.Index.TOKENIZED));
        doc2.add(new Field("content", "ผมหวังว่าตัวอย่างการใช้งานอย่างง่ายนี้ จะเป็นประโยชน์แก่ผู้เริ่มต้น
            ทุกท่านครับ วิชาที่ส จิตกฤตธรรม (ผู้แต่ง)", Field.Store.COMPRESS, Field.Index.TOKENIZED));
        doc2.add(new Field("url", "http://www3.anothersampledoc.co.th/doc2.html",
            Field.Store.YES, Field.Index.NO));

        /**Add the documents to your indexwriter*/
        writer.addDocument(doc1);
        writer.addDocument(doc2);
    }

    public static void main(String[] args) throws Exception {

        ///////////////////////////////////////////////////////////////////
        /////////////////////////////////////////////////////////////////// BEGIN INDEXING PART ///////////////////////////////////////////////////////////////////
        ///////////////////////////////////////////////////////////////////

        /**Let's build the index first.*/
        /**The directory path to store index. This must be a directory (folder).
        * Modify to an appropriate location on your machine.
        */
        String indexPath = "/home/nook/Desktop/demo_index";

        /**Create a Lucene directory.*/
        Directory dir = FSDirectory.getDirectory(indexPath);

        /**An analyzer is used to analyze and tokenize the text added
        to the index. For Thai, use ThaiAnalyzer. You can find it at
        http://sansarn.com/look/download.html*/

        Analyzer analyzer =
        new ThaiAnalyzer();
        // new StandardAnalyzer(); This is for English

        /**Construct an IndexWriter using the index directory and the analyzer.
        The third boolean argument is to specify whether you want to create
        a new index. In this case (true), we will create a new index. If the old
        index exists in the directory, it will be OVERWRITTEN.*/
        IndexWriter writer = new IndexWriter(dir, analyzer, true);

        /**Add some documents to the index so that we can search.*/
        addSampleDocuments(writer);

        /**We optimize the index so that we can search faster.
        * During the indexing, many files (index segments) will be generated.
        This optimization tries to reduces the number of these segment files.*/
        writer.optimize();
    }
}

```

squares

✱ Ake on การเรียนรู้ด้วยวิธี linear least squares

License

Blog under the Creative Commons Attribution-NonCommercial 3.0

License



```

/**DON'T FORGET TO CLOSE THE IndexWriter*/
writer.close();

////////////////////////////////////
//////////////////////////////////// BEGIN SEARCH PART //////////////////////////////////////
////////////////////////////////////
/**Pass the index directory to construct an IndexSearcher*/
IndexSearcher searcher = new IndexSearcher(dir);

/**In Lucene, query is modeled by a Query class. If you are a beginning and don't
want
* to construct the Query object by yourself, you can use QueryParser class.
QueryParser
* will construct a Query object from the query string that you input.
*/
QueryParser qParser = new QueryParser(
"content", //this is the default field to search.
analyzer //The analyzer is needed to tokenize your keywords. For example
//You search โรงเรียนต่างจังหวัด. The analyzer may tokenize the word into
//โรงเรียน and ต่างจังหวัด so that the search can be more flexible.
);

/**Your query*/
String queryString = "ตัวอย่าง วิทย์"; //Normal operator is OR. = ตัวอย่าง OR วิทย์
/**Parse the query and construct a Query object*/
Query luceneQuery = qParser.parse(queryString);

/**Search using the luceneQuery. In Lucene, search results are contained in Hits
object.
One Hit means one hit search result. So, Hits is an object containing a collection
of hit
search results.*/
Hits hits = searcher.search(luceneQuery);

for(int i=0;i
/**This is the document hit by the query*/
Document docI = hits.doc(i);
/**Get the url field*/
String url = docI.get("url");
/**Get the title field*/
String title = docI.get("title");
/**Get the content field*/
String content = docI.get("content");

/**Show all fields to the user*/
System.out.println("URL: "+url);
System.out.println("Title: "+title);
System.out.println("Content: "+content);
System.out.println("-----");
}

/**Don't forget to close the searcher when you are done.*/
searcher.close();

/**Try to change to query to see different results.*/
}
}

```

หวังว่าคงมีประโยชน์แก่ผู้เริ่มต้น 🙏

จำหน่าย พื้นไม้ ประดู่ไม้

สินค้าคุณภาพสูง ในราคาที่ถูกลงกว่า หมดตัวอย่างสินค้าได้ที่
ShowRoom

ลดน้ำหนัก อันดับ 1 ของโลก

หุ่นดีสูตรดารา ไม่โยโย่ อย. 70ประเทศ ไม่ได้ผลคืนเงิน
100%ผู้เชี่ยวชาญดูแล

โฆษณาโดย Google

11 responses so far