Apache Lucene XML file Indexing in C#

Kevin Barthman & Ryan Bemowski

Description

Index XML files with Apache Lucene

Query indexed data

Dynamic determination of XML schema

Use of IKVM for C# GUI

Installation & Execution

Install .Net framework 1.1 or later

Unzip the supplied ‘RyanBemowskiKevinBarthmanProject.zip’ file

Replace files in “xmlDocs” folder (Optional)

Run CIS345FinalApplication.exe

How to Use

Enter query in the “Search Query” textbox.

Optional filter: Choose from “Search Tag”.

Optional filter: Choose from “File to Search”.

Optional filter: Set “Number of Results”.

Execute query by clicking “Run Query”.

Results table will populate if results are found.

Sample Queries

Error handling with improperly formatted XML

All elements and all files, search for “XXX”.
All elements, single file, search for “XXX”.
“YYY” element, single file, search for “XXX”.

Challenges

- How do we index our XML files?
  - We followed the same format as the Apache Lucene demo and researched the documentation

- How do we handle any schema?
  - We read through the XML documents, save all of the XML elements, and index based on the element names

Limitations

No searching of child elements

- Addition of a dictionary of child content

Not aware of children, siblings or ancestors

- Much more complex indexing