LUCON

1. BASIC INFORMATION
    1. Tool name: *LUCON (LUcene-based CONcordancer)*

    2. Overview and purpose of the tool

Lucon allows searching for a single word, for terms, for words at a certain distance. It displays the contexts of the searched terms.

It allows for fast search: while the user introduces characters, the list of words is modified, so that only the matching words are displayed. It can also perform an approximate search: given one word, it displays all words that are close in form to that one. Lucon can search for (compound) terms occurring in a string of words of a length established by the user.

It can also search through the attribute values (only for XML files), which is very useful when you want to search for a lemma (supposing that you have an XML file in which every word is tagged and has a lemma attribute).

Any search can be performed using regular expressions.

Lucon offers the possibility to define the terms (using regular expressions) that should be indexed. For example, in linguistics you are interested to search over a set of words, while in finance you are interested to search over alphanumeric terms (keywords, tickers as S6XOVER or prices as numbers).

It also offers the possibility to select certain XML elements as metadata elements. These are indexed separately from the rest of the elements, but they can be searched for. For example, if you have <ISBN> as an XML element in your document and you do not want its content to appear in the concordance term list, then you can simply add <ISBN> in the list of metadata elements.

The application creates the list of words and their frequency of occurrence in the searched files. These are displayed altogether with a small left and right context of the target word (the user can set the length of this context) and a larger context becomes visible when clicking on the context of interest.

The list of words and their small contexts can be saved in a separate file for further analyses.

    3. A short description of the algorithm

Lucon uses Lucene for indexing and search.

2. TECHNICAL INFORMATION
    1. Software dependencies and system requirements

Lucon is platform independent. It requires Java installed.

    2. Installation

To install the tool, one needs to download an archive from the address http://sourceforge.net/projects/lucon/files/latest/download, extract the files from it and then install it.

    3. Execution instructions

After installation, a corpus needs be indexed and then searches can be made. The indexing is guided and the user can choose the place of the corpus to be indexed, the place for saving the index, the size of the context indexing window, if he wants to index meta-elements in the xml documents. Screen captures at http://sourceforge.net/projects/lucon/ exemplify different usage scenarios for this tool.

4. Input/Output data formats
   a. Input data formats
   UTF-8 text and xml files.
   b. Output data formats
   Concordances in the GUI format.

5. Integration with external tools

Lucon is fully self contained, it depands only on the Java Environment.

3. CONTENT INFORMATION

1. input data
   a test input folder “corpus_test”, containing 23 .xml files
   see test_lucon\corpus_test

2. output data
   some screen captures showing how to use *lucon* with “corpus_test” to extract concordances; *see test_lucon\lucon snapshots*

3. approximation of the time necessary to process the test input data
   30 seconds for the indexing and instant searching;

4. ADMINISTRATIVE INFORMATION

1. Contact
   For further information please contact Cătălin Mititelu:

http://catalinmititelu.users.sourceforge.net/

5. RELEVANT REFERENCES AND OTHER INFORMATION