ASSIGNMENT

Development of a Search Engine and Applications in IR

CI6226 Information Retrieval & Analysis

2015/2016 SEMESTER 2

NANYANG TECHNOLOGICAL UNIVERSITY

1 Objective

The objective of this assignment is to let you getting familiar with the main components in a basic search engine and the applications in Information Retrieval.

2 Assignment Format

1. This is a group assignment. Each group has 4 (at most 5) students.
2. One report is to be submitted by each group and all members in the same group receive the same grade.
3. You may use ANY programming language of your choice, *e.g.,* C, C#, C++, Java, Perl, Python, PHP, Ruby, VisualXXX.
4. You may use a third-party parser to parse the XML file.
5. You must NOT use any relational database (*e.g.,* MySQL) or embedded DB.

3 Dataset

We will use the DBLP data in XML format. DBLP is a computer science bibliography website which indexes more than 2.6 million research papers with attributes like title, author, publication venue and publication time. Please refer to the following pages for more information about the data.

- Dataset in compressed XML and DTD: `http://dblp.uni-trier.de/xml/`
- Dataset FAQ: `http://dblp.uni-trier.de/faq/`
- XML Parser: `http://dblp.uni-trier.de/faq/How+to+parse+dblp+xml.html`
- Schema: `http://dblp.uni-trier.de/faq/What+do+I+find+in+dblp+xml.html`

4 Project

4.1 *Project 1: Develop a Search Engine using Open-Source APIs*

Description: Write a search engine to index and search the publication records listed in DBLP using open-source APIs, *e.g.,* Lucene or other libraries specific to IR.[1] In this assignment, you may use (i) One main IR specific library for most of the operations; (ii) Any other third-party libraries if and only if the main library does not provide the required functionality; and (iii) Any stopword list of your choice.

Detailed Requirements: In DBLP, there are many different kinds of publication records, *e.g.,* article published in a journal or magazine, research paper published in a conference, book, and PhD thesis. We are interested in two kinds of publication records only:

[1] `http://en.wikipedia.org/wiki/List_of_information_retrieval_libraries`

- **article** — An article from a journal or magazine.
- **inproceedings** — A paper in a conference or workshop proceedings.

For article and/or conference paper, you need to index at least the following attributes: paper id (or key), paper title, authors, year of publication, and the publication venue (*e.g.,* ⟨journal⟩ for ⟨article⟩ and ⟨booktitle⟩ for ⟨inproceedings⟩).

- *Indexing*. You should use the library APIs to evaluate the impact of (i) performing or ignoring stemming on the title attribute of the publication records, (ii) distinguishing or ignoring upper/lower case characters on the title attribute, and (iii) using or not using stopwords on the title attribute. The evaluation shall include the change in the speed of indexing, the size of vocabulary on the title attribute.
- *Queries*. Your system should support (i) free text keyword queries on any attributes, and (ii) free text keyword queries on specific attributes (*e.g.,* title, author, publication venue, and/or the combination of attributes). In free text keyword queries, a phrase can be specified by using double quotation marks, *e.g.,* “event detection”.
- *Query Results*. Top N (the number of N is configurable) results should be returned via the console/screen along with rank, scores, docID, and snippets whenever possible.
- *Evaluation*. You are required to evaluate the search accuracy of your search engine. The evaluation shall report (*whenever possible*) Precision, Recall, and F1, for the top-10 results of 10 sample queries selected at your own choices. You may use these queries to demonstrate that your search engine is capable of processing the different kinds of queries listed above.
- *User Interface*. A text-based command line system is sufficient. A GUI or web-based interface to the search engine is encouraged.

4.2 *Project 2: Develop two IR Applications*

- List the top-10 most popular research topics in each year from 2000 to 2015. The research topics are to be extracted from the titles of the publication records. A research topic can be a single word (*e.g.,* optimization, recommendation), or a phrase (*e.g.,* “topic model”, “event detection”). You may further extend your application to list the top-10 most popular research topics in each year for a specific conference, journal, or an author.
- If we consider all the paper titles published in a single year in a publication venue as a virtual document (*e.g.,* all papers published in SIGIR 2015, or all the papers published in the IEEE Transactions on Data and Knowledge Engineering (TKDE) in 2014), then we can search for the most similar publication venue and year, for a given conference or journal as a query (*e.g.,* CIKM 2014, TKDE 2012 are two example queries). List the top-10 most similar publication venue and year for at least 6 sample queries of your choice.

You may reuse the indexes constructed in Project 1 or re-index the publication records to support the applications in Project 2.

5 Submission of Report and Source Code

5.1 *Hardcopy Report*

- The hardcopy report must be submitted on or before **04 Apr 2014** (Monday, Week 12). The report shall be formatted following the ACM SIG Proceedings Templates (either MS Word or Latex), maximum 8 pages including appendix if any.[2] DO NOT include in your report all the source code and complete results sets. However, you must include *code snippets* which are important for the main functions of your system. You should cite all third-part libraries used in your assignment.

- The report shall be printed in double-sided format whenever possible. A plastic cover or ring-binding leads to 2% penalty.

5.2 *Source code, documentation, and softcopy report*

- A CI6226.zip file containing the following files and folder shall be submitted: Report.PDF, Readme.txt, SourceCode.

  - Report.PDF shall be the same as the hardcopy report submitted.
  - Readme.txt shall include a link to download the third-party library if you used any in your assignment, an installation guide on how to setup your system assuming the dataset is provided, how to use your system (*e.g.,* command lines, input format, parameters) and explanations of sample output obtained. Readme.txt shall also include sample output for example queries used in your report.
  - SourceCode folder shall contain all your source code. The dataset and the indexes produced by your code shall **NOT** be included in the softcopy submission to minimize the file size.

- Softcopy submission deadline: ***04 Apr 2014 (Monday) 11:59PM***. Late submissions are allowed but will be penalized by 0.5% every calendar day (until zero). The softcopy can be submitted for at most three times, only the last submission will be graded and time-stamped.

[2] `http://www.acm.org/sigs/publications/proceedings-templates`

Paper Presentation

The paper presentation is tentatively scheduled on Week 11 and Week 12. At least two students from each group will represent the group to present the paper chosen by the group members. Each presentation lasts for at most 20 minutes followed by 3-5 minutes Q&A. A random generator (which generates numbers from 1 to 5 randomly) will be used to choose which member (M1-M5) in a group to present the first 10 minutes. The group may recommend another member to present the remaining slides. Questions may be answered by any member in the group.

The papers to be presented must be a full research paper (8 - 10 pages) chosen from one of the following conferences: SIGIR 2013 – 2015, WSDM 2013 – 2015. You may follow these links to check the paper titles. The PDF of your selected paper can be downloaded from ACM Digital Library through NTU library databases. Once a paper is selected, you should update the Google Doc so that other groups cannot choose to present the same paper. All groups are required to select the paper and update the online document by 18 Mar 2015.

- `http://dblp.uni-trier.de/db/conf/sigir/sigir2013.html`
- `http://dblp.uni-trier.de/db/conf/sigir/sigir2014.html`
- `http://dblp.uni-trier.de/db/conf/sigir/sigir2015.html`
- `http://dblp.uni-trier.de/db/conf/wsdm/wsdm2013.html`
- `http://dblp.uni-trier.de/db/conf/wsdm/wsdm2014.html`
- `http://dblp.uni-trier.de/db/conf/wsdm/wsdm2015.html`