Data Management
Lab2 : Evaluation of results obtained from Lucene

Prof. Nastaran FATEMI
FATEMEH BORRAN

1. Objectives

The goal of this lab is to evaluate the similarity function implemented in Lucene using different analyzers and by calculating a set of the metrics introduced in our course.
We will use the same corpus as the first lab: the CACM publication list.

Organization: The lab should be realised in a group of maximum 2 students.

Report: You will return a report containing both the answers to the questions asked in the different sections, and the source code of your implementation.

Deadline: Friday 20.11.2015 before midnight on Cyberlearn:
https://cyberlearn.hes-so.ch/mod/assign/view.php?id=514522

2. How to proceed

In this lab, you will make a comparative evaluation of using four different analyzers:

1- WhiteSpaceAnalyzer
2- StandardAnalyzer
3- EnglishAnalyzer with default stopwords, and
4- EnglishAnalyzer with common_words.txt

You will apply one-by-one each of the four mentioned analyzers on the CACM collection and compare the results.

To be able to evaluate results you need a benchmark. The CACM collection provides you such benchmark. It provides you a set of **queries** and their respective **relevant results.**

CACM Queries
The CACM collection contains 64 queries given in query.txt. The file contains the query id and the query text separated by tabulation. There is an empty line between each query.

CACM List of relevant results
The CACM collection has also a list of relevant results per query provided in qrels.txt, which correspond to the 64 queries. For example query 1 has 5 relevant results that include publications 1410, 1572, 1605, 2020 and 2358. Note that there are queries with no relevant documents (for example queries 50-56).

Data Management
Lab2 : Evaluation of results obtained from Lucene

Prof. Nastaran FATEMI
FATEMEH BORRAN

3. Steps of evaluation

I. Indexing

- Create an index with two fields: publication id and publication content. Publication content should contain both the title and the summary of the publication.

- Make four different indexes each time using one of the four analyzers : WhiteSpaceAnalyzer, StandardAnalyzer, EnglishAnalyzer with default stopwords and EnglishAnalyzer with common_words.txt.

II. Quering

For each query, use the query parser to search the query and note all the corresponding results. Use QueryParser.escape(queryString), where queryString is the query text in String, before parsing the query to escape all non-allowed characters of the query by the QueryParser.

III. Evaluation

Compare the following metrics:

1) **Summary Statistics:**
   a. total number of documents,
   b. total number retrieved documents for all queries
   c. total number of relevant documents for all queries and
   d. total number of relevant documents retrieved for all queries.

2) **Average Precision at Standard Recall Levels:** Precision at 11 standard recall levels (0, 0.1, 0.2, …, 0.9, 1.0). Each precision average is computed by summing the interpolated precisions at the specified recall value and then dividing by the number of queries.

3) **Recall-Precision Graph:** It is created using the 11 recall values obtained in the last step. You can use Excel to represent the graph. Show one curve per analyzer in the graph.

4) **Average precision over all relevant documents (MAP):** As an example, consider a query that has four relevant documents which are retrieved at ranks 1, 2, 4, and 7. The actual precision obtained when each relevant document is retrieved is 1, 1, 0.75, and 0.57, respectively, the mean of which is 0.83. Thus, the average precision over all relevant documents for this query is 0.83.

5) **Average R-Precision:** R-Precision is the precision after R documents have been retrieved, where R is the number of relevant documents for the query. The average R-Precision is computed by taking the mean of the R-Precisions of the individual queries.

IV. Reporting the results

- Make tables to compare the numeric results and visualize the graphs recall-precision. Make a conclusion on your results.