Information Retrieval

Practical session n°5: Structured IR at INEX

Create a directory named *practice5*. In this directory, create a new file named *practice5_report.txt*.
For each exercise, copy-paste in this file some outputs of your program, showing that you complete the exercise and it works correctly. Add some explanations.

You will build several runs (maximum = 50 runs / submission) for INEX Ad-Hoc Relevant in Context task. Each run is stored in a textual file, which should be named using the following template (see Practical session n°4 for details): *TeamName_Run-Id_WeigthingFunction_Granularity_Stop_Stem_Parameters.txt*

Example: VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt

At the end of your work:
- copy-paste the source code of your program(s) in the directory *practice5*.
- copy all your runs (maximum = 50 runs) in the directory *practice5* (without any sub-directory).
- compress the directory *practice5* in a file named *practice5_YourTeamName.zip* (or *.tar*, *.gz*, *.rar*, etc.) (e.g.: *practice5_VictorAlbertJulesIsaac.zip*).
- upload this compressed file (one file / team) on the website of the course **before 11.59pm November 21st**.

The week after, repeat the whole process: create a directory named *practice5v2*, a file *practice5v2_report.txt*, copy your source code, copy all your runs (maximum = 50 runs), compress the directory *practice5v2* in a file named *practice5v2_YourTeamName.zip* (or *.tar*, *.gz*, *.rar*, etc.) and upload this compressed file (one file / team) on the website of the course **before 11.59pm December 5th**.

And again, repeat the whole process: *practice5v3, practice5v3_report.txt, practice5v3_YourTeamName.zip* … and upload this compressed file **before 11.59pm December 12th**.

And again, repeat the whole process: *practice5v4, practice5v4_report.txt, practice5v4_YourTeamName.zip* … and upload this compressed file **before 11.59pm January 3rd,**.

Exercise 1: Indexing XML documents

Download the collection of XML documents *Practice_05_data.zip* on the website of the course.
It contains 9,804 XML documents stored in 9,804 XML files (237MB).
Index this collection, without considering XML tags, and using your indexing program and the weighting function of your choice.
Compute again the collection statistics and check if they are the same as those computed during the practical session n°3.

Exercise 2: SMART *ltn* run (XML elements)

Retrieve a ranked list of 1,500 XML elements from the collection of the Exercise 1, for each of the 7 queries of the Practical session n°4, using SMART *ltn* weighting function, using stop-words or not, and using a stemmer or not. Fix the granularity (i.e. list of XML tags considering as potential indexing unit) as you want.
With these results, build one or several runs for the RIC task of INEX. Check carefully the syntax.

Exercise 3: SMART *ltc* run (XML elements)

Same question, but using SMART *ltc* weighting function instead of *ltn*.

Exercise 4: *BM25* run (XML elements)

Same question, but using *BM25* weighting function instead of *ltc*. Set the k_1 and b parameters as you want. You can use these usual values: $k_1 = 1.2$ and $b = 0.75$.

Exercise 5: fields weighting [Wilkinson94]

Retrieve a ranked list of 1,500 documents, for each of the 7 queries, using the late combination of fields proposed by [Wilkinson94], and using *BM25* weighting function. Fix the granularity (i.e. list of XML tags considering as fields). Set the k_1, b, and α_i parameters as you want.
With these results, build one or several runs for the RIC task of INEX.

Exercise 6: fields weighting [Robertson94]

Same question, using the early combination of fields proposed by [Robertson94].