TERM PROJECT

406.424 Internet Applications, Spring 2010

Term Project Description:

Develop a scoring function that can rank documents against a query according to their relevance by utilizing the following information available from Galago toolkit.

- Corpus length
- The number of documents in the corpus
- Document length
- Document frequency
- Term frequency per document

You are asked to enhance one of the retrieval models taught in the class, and implement your scoring function in Java using Galago toolkit. Performance of your scoring function will be evaluated by taking the performance of BM25 model as a baseline. More specifically, the evaluation metric will be Mean Average Precision (MAP) @ K, where K = 10, and your performance score will be determined by the following formula, where MAP_{base} represents MAP@10 performance of the baseline model and MAP_{prop} indicates MAP@10 of the model you propose.

$$score = \begin{cases} 100 & \text{if } 0.5 \le MAP_{prop} \\ \dfrac{80}{1-2\cdot MAP_{base}}(2\cdot MAP_{prop}-1)+100 & \text{if } MAP_{base} < MAP_{prop} < 0.5 \\ 20 & \text{if } MAP_{prop} \le MAP_{base} \end{cases}$$

To help your work, the following materials will be provided from the course home page (Note: You should write your Java code in IntappScorer.java file and should not change the provided directory structure of the Galago toolkit):

- Galago toolkit
- IntappScorer.java file: empty java file prepared for your implementation
- Example Java files for the query likelihood ranking model and BM25
- Document collection: cacm.corpus
- Performance evaluation method

As a project output, you are supposed to submit the followings:

- **Report** that contains detailed descriptions of your proposed retrieval model and performance evaluation results
- **Presentation material** for your achievements
- **IntappScorer.java** file you implemented for your scoring function

The Guidelines:

- The term project of this course is *team-based*. Each team may consist of up to **2 students**. ~~It is required to report the team composition via email (to the instructor cc TA) by March~~.

- Each team is expected to submit a term project report as well as presentation material, which are due by June 10th, 2010.

- All the submissions should be made in the forms of *email* as well as *hard copy* to the instructor and TA. While the hard copy submission should be made in-class on the presentation day (i.e., June 10th), email submission is due by the midnight before that day. It is important to note that **you are not allowed to change any part of the materials after submission**.

- We will have presentation sessions on June 10th, and each presentation will be 10 min. long.

- There will be small additional bonus points if the report is written in english or the presentation is given in english.

The Grading:

- The performance score of your retrieval model: 70%

- Underlying rationale and justification on your retrieval model: 20%

- Report quality: 10%