King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering
CPE455/671 Search Engine and Internet Mining, 2/2014

Assignment 2: Probabilistic retrieval
Due: in 4 weeks, 22 April 2015, 23:59

Instruction

This assignment is based on what you have developed in Assignment 1. Ten text document files, which are from Wikipedia of 10 movies, are given. Assume that the relevant movies that we look for are only Gravity and Interstellar.

1. Create inverted indices of these documents using global position. (similar to Assignment 1)
2. Implement a lookup function, **getDocID**, which return the document id number. (Hint: you will need a data structure with three column {doc_id, start_pos, end_pos})
3. Compute the Robertson/Sparck Jones weight w_t of all the weight t in all the 10 documents using Eq. 6 in the lecture.
4. Implement a query search function that prints the ranked documents based on their combined weights.

Grading

The assignment will be graded based on the best level of achievement you can accomplish.

For further questions, please ask me in the class or via Facebook group.

Submission

You have a choice to submit just some of them if you cannot do the whole assignment, but your assignment score will be deducted appropriately.

a) A text file contain the inverted indices (result of Task 1)
b) A text file contain a lookup data structure (result of Task 2)
c) A text file containing two columns {term t, w_t} (result of Task 3)
d) Test results: Search queries a test results (result of Task 4)
e) Source codes of the above

Students can choose to either (1) submit the assignment via e-mail to santitham@cpe.kmutt.ac.th or (2) submit in person by making an appointment and bringing all the files to me.

Academic misconduct

All types of academic misconducts are prohibited. Finding such action will result in a serious academic consequence.