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

Assignment 1: Basic technique in search engine
Due: in 4 weeks, 19 February 2015, 23:59)

Instruction

Given the text file "frogprince.txt", perform the following tasks

1. Normalize the text by (1) convert all letters to lowercase letters
2. Remove other symbols, for examples ".", ":", "!", and others
3. Tokenize the normalized text into tokens and their location. Print out some location on the text file to show that it works.
4. Implement the in-memory array-based version of inverted indices presented in the lecture using the binary search of the **next** method. Show that the **next** method works.
5. Use your implementation of inverted indices, implement the basic phrase search algorithm presented in the lecture. Show that it works by printing out some results.

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 all lower case letters (result of Task 1)
b) A normalized text file without any symbols (result of Task 2)
c) List of all tokens together with their positions (result of Task 3)
d) A text file showing the inverted indices (result of Task 4)
e) A source code for phrase search and a test result of specific phrase search (result of Task 5)

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.