Concordia University
Department of Computer Science and Software Engineering
COMP474/6744 Assignment 3 Fall 2007

November 14, 2007

Due Date: December 11, 2007

Exercises:

1. (3 pts) Word frequency
   (a) (1/2 page) Describe Zipf's law and how it applies to frequency of words in English. Give some data that you retieve from the Web (the British National Corpus may be a good source, many statistics are already precompiled and have been traded on several mailing lists).
   (b) (1/2 page) Describe what is meant by "the long tail" and compare the usage of the term to the usage of the term "Zipf's law".
   (c) (1 page) Discuss what the implications for IR strategies are when considering the long tail as a valuable part of search space. Discuss in particular the suitability of tf-idf ranking for keywords from the long tail.
2. (3 pts) Indexing
   Using the .html files from http://www-cse.ucsd.edu/ rik/foa/l2h/, sections 1,2,3,4,and 6, compile an index for all words using UNIX shell commands or your favorite scripting language.
   (a) What is its size? How does its size compare to the index of the book?
   (b) Reduce the index to a more useful size. Which terms do you suppress? Why? What is the resulting size of your index? You may experiment with different strategies and compare the results. Submit and compare your favorite index with the official book index.
3. (4 pts) The Web as a Super Expert System
   Give a critical discussion of the following paragraph taken from Belew: Finding Out About, Section 6.9 (limit yourself to three pages, give a scolarly discussion using the notions and terms introduced throughout class and drawing on your experience with writing a small expert system):

   > Note the ease with which author-as-knowledge engineer can express their knowledge. Hypertext knowledge bases are accessible to every writer. In this view, hypertext solves the key AI problem of the KNOWLEDGE ACQUISITION BOTTLENECK, providing a knowledge representation language with the ease, flexibility and expressiveness of natural language — by actually using natural language! The cost paid is the weakness of the inferences that can be made from a textual foundation: contrast the strong theorem-proving notions of inference of Section 6.5.1 with the many confounded associations which arise in Swanson's analysis of latent knowledge in Section 6.5.3 .
4. (0 pts) Feedback question: has your search behavior changed during the assignment or as a result of class? If so, in what way?

Requirements:

- This assignment has to be done individually. You may discuss the topic with your fellow students in general, but the write-up has to be done individually.
- The assignment has to be submitted electronically as "assignment 3" at:
  `https://eas.encs.concordia.ca/eas/authentication.jsp`
- The Expectations of Originality form has to be submitted.