CS4611 Information Retrieval 07/12/2012

Assignment 3: Review 2

Professor: Michel Schellekens *TA:Ang Gao*

Assigned: December 7, 2012 Due on: December 14, 2012 in class

Notice: Please submit your assignment on December 14 2012 during tutorial time and write down your name and student ID .

Problem 1.
Write down VB code and Gamma code for number 777. **[5 points]**

Problem 2.

- Looking at a collection of web pages, you find that there are 7000 different terms in the first 20,000 tokens and 36,000 different terms in the first 1,500,000 tokens.
- Assume a search engine indexes a total of 30,000,000,000 (3×10^{10}) pages, containing 200 tokens on average
- What is the size of the vocabulary of the indexed collection as predicted by Heaps' law (log based 10 and round result to integer)?

[7 points]

Problem 3.
What is the Levenstein distance between the following pairs of strings? "thorough and "throughout, write down your calculation steps. **[7 points]**

Problem 4.
Define the terms recall and precision. **[5 points]**

Problem 5.
The F-measure is defined as the harmonic mean of recall and precision. What is the advantage of using the harmonic mean when compared to the arithmetic mean? **[5 points]**

Problem 6.
Consider a web graph with three nodes 1, 2 and 3. The links are as follows: $1 \rightarrow 2$, $3 \rightarrow 2$, $2 \rightarrow 1$, $2 \rightarrow 3$. Write down the transition probability matrices for the surfers walk with teleporting, for the following three values of the teleport probability: (a) $\alpha = 0$; (b) $\alpha = 0.5$ and (c) $\alpha = 1$. **[6 points]**

Problem 7.
Show that the PageRank of every page is at least α/N. What does this imply about the difference in PageRank values (over the various pages) as α(teleporting probability) becomes close to 1? **[5 points]**