CS4611 Information Retrieval 11/23/2012

Assignment 2: Review 1

Professor: Michel Schellekens *TA:Ang Gao*

Assigned: November 23, 2012 Due on: November 30, 2012 in class

Notice: Please submit your assignment on Nov 30 2012 during tutorial time, and write down your name and student ID .

Problem 1.
Consider these documents:

- Doc1: solution found for laziness
- Doc2: old laziness found
- Doc3: old approach for treatment of laziness
- Doc4: old hopes for laziness patients

A: Draw the term–document incidence matrix for this document collection.
B: Draw the inverted index representation for this collections. **[10 points]**

Problem 2.
Recommend a query processing order for the following Boolean query:
(bush OR apricot) AND (pudding OR brown) AND (phones OR ears)
Assume that the document frequencies of the terms in the above query are:

- ears 11331
- phones 9700
- pudding 10091
- brown 27160
- bush 4660
- apricot 21681

Explain the rationale of the order that you found. **[5 points]**

Problem 3.
Why are skip pointers not useful for queries of the form x OR y? **[5 points]**

Problem 4.
Write down the entries in the permuterm index dictionary that are generated by the term **cork** **[5 points]**

Problem 5.
How do stopping and stemming reduce the size of an inverted index? **[5 points]**

Problem 6.
Consider the following collection:

- Doc1: new york times
- Doc2: new york post
- Doc3: los angeles times

Given the query **new times**, using tf/idf ranking documents.
Notice: tf–weight using: $1 + \log \mathrm{tf}_{t,d}$ and *log* is base 10. **[10 points]**