Information Retrieval and Information Extraction

First Mid-Semester Examination, Spring 2016

Maximum Marks: 90 *Maximum Time: 90 minutes*
Answer all questions and keep your answers short and to the point

1. Draw the inverted index that would be built for the following document collection.

    a. Doc 1 new home sales top forecasts
    b. Doc 2 home sales rise in july
    c. Doc 3 increase in home sales in july
    d. Doc 4 july new home sales rise

(10 Marks)

2. Recommend a query processing order for the query: (tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes) given the following postings list sizes:

Term	Postings-size
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

(10 Marks)

3. What is a relevance function? Provide one relevance function each for each class of the IR models. (10 Marks)

4. What is the typical ratio between the sizes of the document collection and the index? Is it the same for small collection and a very large collection? (10 Marks)

5. Will this ratio be different between languages? For example, will the index size of be different for 10,00,000 words in 10,000 documents of English and same number of documents in Telugu, Hindi, German and Finnish? Justify your answer with details on Index scheme. (15 Marks)

6. Estimate the space usage of the Reuters-RCV1 dictionary with blocks of size $k = 8$ and $k = 16$ in blocked dictionary storage. Assume 400000 words, 8 bytes per word, 4 bytes to store frequency and 4 bytes for postings pointer. (15 Marks)

7. What is the best way to efficiently incorporate static quality scores for documents in inverted indexes? (10 Marks)

8. What are the three main issues with Jaccard similarity for scoring a document wrt a query? (10 Marks)