Information Retrieval Supervision 1 (2017/18)

Question 1

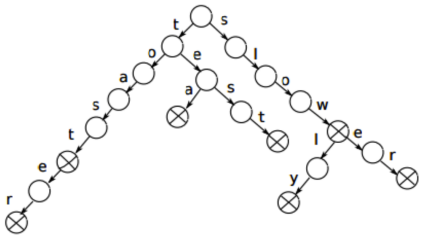
1. Draw the term-document incidence matrix and the inverted index
   - Doc 1: breakthrough drug for schizophrenia
   - Doc 2: new schizophrenia drug
   - Doc 3: new approach for treatment of schizophrenia
   - Doc 4: new hopes for schizophrenia patients
2. Using the document collections above, what will be returned for the following queries
   - Query 1: schizophrenia AND drug
   - Query 2: for AND NOT (drug OR approach)

Question 2

1. Which of the following statements are true?
   - In a Boolean retrieval system, stemming never lowers precision
   - In a Boolean retrieval system, stemming never lowers recall
   - Stemming increases the size of the vocabulary
   - Stemming should be invoked at indexing time but not while processing a query

Question 3

1. In the inverted index of an information retrieval system, dictionary terms can be represented using different data structures.



   A Consider the trie in the figure above, which encodes several dictionary terms.
      i. List the terms contained in this trie
      ii. Explain how terms are looked up in a trie

B Alternatively, we could store the terms in a binary search tree

i. Draw the binary search tree with minimal depth that stores the dictionary terms from the figure above.

ii. Compare the worst-case time complexity of the dictionary lookup for a binary tree and a trie. What are the conditions where the binary tree is preferable to a trie.

C Next consider a radix tree, a spaced-optimised trie data structure where each node with only one

i. Draw the radix tree containing the dictionary terms from the figure above.

ii. Give an algorithm for insertion of a new index term $t = t(0)...t(k)$ into a radix tree. Use examples to illustrate your algorithms. You may use pseudocode as long as you can clearly explain your thoughts.

Question 4

1. Compute the Jaccard matching score and the tf matching score for the following query-document pairs.

query	document
information on cars	all you've ever wanted to know about cars
information on cars	information on trucks, information on planes, information on trains
red cars and red trucks	cops stop red cars more often

Question 5

1. Assume you have to explain the tf-idf weighting to someone unfamiliar with the scheme. In one or two paragraphs, explain the intuition behind the tf-idf weighting scheme.

2. How does the tf-idf scheme exploit the phenomenon known as Zipf's law for assigning weights?