資訊網路與多媒體研究所

資格考試科目：資訊檢索與擷取

1. (20 pts) Consider a query q for which there are 6 relevant documents in the collection of 100 documents ($d_1 \ldots d_{100}$). Let (d_2, 1), (d_{12}, 3), (d_{25}, 3), (d_{32}, 2), (d_{54}, 1), and (d_{60}, 2) be the pairs of (document ID, relevance score) for the 6 relevant documents. All other documents are judged as irrelevant. Contrast two systems run on this collection. The ranked lists of their search results are shown as follows:

   System 1: $d_{54}, d_{40}, d_{12}, d_{70}, d_{32}, d_{18}, d_{86}, d_{72}$

   System 2: $d_{30}, d_{59}, d_{25}, d_{60}, d_{17}$

   Answer the following questions. Show your calculation in (a) and (c).

   (a) (4 pts) Which system has a higher precision? Which system has a higher recall?

   (b) (5 pts) For the same system, is there always a trade-off between precision and recall? Explain your answer.

   (c) (6 pts) What is the NDCG of system 1? What is the MAP of system 2?

   (d) (5 pts) What are the similarity and difference between MAP and NDCG?

2. (20 pts) The probability ranking principle provides a theoretical basis for probabilistic retrieval models. Binary Independence Model (BIM) and Okapi BM25 weighting scheme are two famous ways of building such probabilistic models.

   (a) (4 pts) What is probability ranking principle?

   (b) (4 pts) Give two examples to explain why search results that are optimal according to the probability ranking principle are not always ideal from a user's perspective.

   (c) (8 pts) Discuss in details about the difference between BIM and Okapi BM25 in terms of term frequency, inverse document frequency, and document length normalization.

   (d) (4 pts) Describe how to apply relevance feedback in BM25.

3. (20 pts) Language modeling (LM) is a quite general formal approach to IR.

   (a) (4 pts) Given a document d containing m words, i.e., $w_1 \ldots w_m$, describe how to calculate the probability of generating document d based on an n-gram language model θ.

   (b) (4 pts) Give two examples to explain why Dirichlet Prior smoothing is preferred than Laplace smoothing in IR.

   (c) (12 pts) There are three ways to think of using LM in IR, including query likelihood model, document likelihood model, and KL-divergence-based model comparison. Please describe each of the models based on their formulas and then compare their advantages and disadvantages.

4. (20 pts) Compression for inverted index is essential for an efficient IR system.
   (a) (4 pts) Describe how to effectively compress the postings list for stop words.
   (b) (5 pts) What is the postings list that can be decoded from the following variable byte-code?

       10001001 00000001 10000010 11111111

       What would be the encoding of the same postings list using a γ-code?
   (c) (6 pts) Consider a two-word query. For one word, the postings list consists of the following 16 entries:

       [4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 122, 157, 180].

       For the other, it is the one entry list: [47]. How many comparisons would be done to intersect the two lists if the postings lists have skip pointers with a skip length of 4. Show your calculation. Discuss if a larger skip length always reduces the number of the comparisons. Explain your answer.
   (d) (5 pts) Discuss how the presence of skip pointers affects the choice of compression methods.

5. (20 pts) There are lots of question-answer pairs available on online forums. Please develop a forum retrieval system. Given a query, the system is expected to rank the answers according to their relevance and quality. Describe what resources required, your algorithms (show the formulas), and how to conduct experiments to evaluate your results. Be specific as you can.