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. (22 pts) Assume one query has 6 relevant documents in the collection of 50 documents. Below is a table showing how the 6 relevant documents are rated. 3 means high relevance while 1 means low relevance. All of the other documents, i.e., those not shown in the table, are judged as irrelevant.

document ID	d_1	d_2	d_9	d_{17}	d_{25}	d_{42}
relevance score	3	1	2	3	2	1

Two IR systems rank results with respect to the query as follows (the leftmost document is the top ranked one):

System 1: d_1 d_{25} d_7 d_{21} d_2 d_{39}

System 2: d_9 d_{21} d_6 d_2 d_{19}

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

(a) (5 pts) What is the MAP of System 1?
(b) (5 pts) What is the NDCG of System 2?
(c) (6 pts) Calculate the kappa measure for agreement between the two systems.
(d) (6 pts) Discuss the relationship between recall and MAP@k.

2. (20 pts) The basis of probability ranking principle (PRP) indicates that a retrieval system performs optimally if documents are ranked according to decreasing probabilities of their relevance to a query. The binary independence model (BIM) is developed based on PRP.

(a) (8 pts) Describe how BIM with relevance feedback estimates the probability that a word appears in a document relevant or irrelevant to a query.
(b) (6 pts) Give an example to explain that ranking based on PRP may not provide satisfactory search results even if the probability of relevance can be well estimated.
(c) (6 pts) Give two advantages of TF-IDF-based vector space model over BIM.

3. (28 pts) Inverted index can improve the speed of query processing in IR systems.

(a) (6 pts) Give two advantages of controlled-vocabulary indexing over free-text indexing.
(b) (6 pts) What are term-at-a-time and document-at-a-time query processing methods, respectively? Give an advantage of term-at-a-time query processing over document-at-a-time query processing.
(c) (8 pts) Give an example to describe how to effectively compress a postings list, and then discuss how document frequency of a word affects the compression ratio of its posting list.
(d) (8 pts) Write an algorithm to efficiently intersect two given postings lists with skip pointers, and then discuss how the number of skip pointers affects the complexity of your algorithm.

4. (20 pts) Statistical language models (LM) have been successfully applied to IR.
   (a) (10 pts) What are query likelihood LM and document likelihood LM, respectively? Give an advantage of query likelihood LM over document likelihood LM in IR.
   (b) (10 pts) Propose a way to perform smoothing for a LM (show your calculation), and then discuss how the smoothing affects retrieval performance for (1) the query terms not occurring in a document and (2) the query terms occurring in a document.

5. (10 pts) More and more review comments are generated on the Web. Please develop a method to perform sentiment analysis, whose goal is to determine if the expressed opinion in a given document is positive, negative, or neutral. Describe how to extract sentiment words as features, how to determine polarities, and how to evaluate the proposed method. Discuss what difficulties you might face in this task. Be specific as you can.