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. (12 pts) Given a query, an IR system returns a ranked list of 10 documents. Assume there are 5 relevant documents for this query, and the documents ranked in the 1st, 2nd, 4th, and 8th positions are relevant. Calculate the 11-point interpolated precision values at the recall levels of 0.0, 0.1, 0.2, …, 1.0 for this query.

2. (28 pts) It is common to represent documents as vectors in a vector space for text classification.
   (a) (9 pts) What are the three major term weighting heuristics in the vector space model and how are they calculated?
   (b) (8 pts) Describe how Rocchio and kNN algorithms perform classification, respectively.
   (c) (5 pts) A multimodal class is the class that has multiple clusters in the vector space. Please explain why kNN handles multimodal classes better than Rocchio.
   (d) (6 pts) Specify leave-one-out cross-validation. When will you use it to evaluate classification performance?

3. (20 pts) Smoothing is a key issue for both of the query likelihood language model and naïve bayes classification model.
   (a) (5 pts) What happens if we adopt language model or naïve bayes model with no smoothing in IR?
   (b) (15 pts) Take language model or naïve bayes model as an example. Give one smoothing method (show your formula). Then analyze its advantages and disadvantages.

4. (10 pts) Both language modeling and probabilistic retrieval model operate in terms of probabilities. What are their differences in retrieval and relevance feedback? Discuss your answers based on their formulas.

5. (15 pts) Consider the simple web graph with 3 pages (A, B, and C), where page A points to pages B and C, page B points to page C, and page C points to page B.
   (a) (5 pts) Assume the probability that a person randomly chooses an out-link to follow at any step, i.e., the damping factor, is 0.9. Compute the transition probability matrix underlying the graph.
   (b) (10 pts) Compute the PageRank score for each of the three pages. Show your calculation.

6. (15 pts) Propose a relevance feedback method for content-based image retrieval (CBIR). When querying images by an example image, one can indicate whether or not the images initially returned from a CBIR system are relevant. Your method should learn how to improve search results based on such feedback information. Describe what features you use, how your method works, and what evaluation metric(s) you adopt to evaluate if the retrieval performance is improved accordingly.