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) Several variants of term-weighting for vector space model have been developed.
   (a) (6 pts) The logarithm function is often used for calculating some weights. Give one example formula for such weight. Explain the rationale behind the usage of logarithm as clearly as possible.
   (b) (10 pts) Here is the way to transform term frequency (TF) in Okapi BM25:

   $$\frac{(k+1)\cdot TF}{k+TF}$$ (k is a non-negative number)

   What's the meaning of parameter k? Discuss the cases where $k = 0$ and $k = \infty$.
   What's the upper bound of the transformed TF?
   Draw a figure to show the relationship between original TF and transformed TF.
   (c) (4 pts) We want to rank documents $\{d_i\}$ (i=1..N) w.r.t. query q. Prove that if q and d_i are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.

2. (16 pts) NDCG (Normalized Discounted Cumulative Gain), MAP (Mean Average Precision), Precision@k and F-measure are famous evaluation metrics in IR.
   (a) (6 pts) What are the important aspects of relevance that are considered by NDCG but not by F-measure?
   (b) (6 pts) Compare the performance of two IR systems. Give an example to explain if it's possible that one performs better in terms of MAP but worse in terms of Precision@k? Show your calculation.
   (c) (4 pts) Consider the balanced F-measure. What is the advantage of using the harmonic mean rather than arithmetic or geometric mean in its formula?

3. (28 pts) Most of search engines extract two kinds of features from web pages and structures, including topical features (e.g., terms) and quality features (e.g., PageRank), and apply machine learning approaches to learning the ranking.
   (a) (8 pts) Describe four lexical processing methods for topical features.
   (b) (9 pts) Suppose a transition probability matrix defines a Markov chain. Under what conditions will the PageRank scores reach a stationary distribution?
   (c) (6 pts) Supervised learning requires training data to learn a model. How do search engines obtain such training data in an automatic way? Give an example to demonstrate how to learn a model for determining the parameters that combine topical and quality features.

(d) (5 pts) How do search engines test if a new model is better than an old one? Describe how to reasonably refine an online IR system with a large number of users.

4. (16 pts) Language modeling (LM) has been well applied to IR. Smoothing always plays an important role in LM. Given document d, word w, and reference corpus C, please answer the following questions.
   (a) (5 pts) Describe how to apply LM to IR.
   (b) (5 pts) A common solution to the smoothing problem is to use a reference model for unseen words. In this case, $P(w|d)$ would be
       (1) $P_{seen}(w|d)$ if w appears in d, or
       (2) $\alpha\ P(w|C)$ if w isn't in d, where α denotes a weight.
       Suppose $P_{seen}(w|d)$ and $P(w|C)$ are both given. Please determine α based on $P_{seen}(w|d)$ and $P(w|C)$. Show your calculation.
   (c) (6 pts) A more general form of $P(w|d)$ is a mixture model defined as follows:
       $P(w|d) = \lambda\ P(w|d) + (1-\lambda)\ P(w|C)$ (λ is a weight varying from 0 to 1)
       Please show how to derive parameter λ according to the EM algorithm.

5. (20 pts) Many clustering algorithms need the number of clusters to be given in advance such as parameter k for k-means.
   (a) (10 pts) Suppose k-means defines its cost function to minimize average (squared Euclidean) distance between each data point and its closest centroid. Prove the convergence of k-means.
   (b) (5 pts) Explain why k-means converges to a local optimum.
   (c) (5 pts) Give an automatic way to determine k. Explain your answers.