Information Retrieval and Extraction
Midterm Examination (2017.11.9)

1. Index terms can be used to represent document contents for effective and efficient retrieval. Please describe how to select "good" index terms from a document. (10 points)

2. Capturing term dependencies is a way beyond traditional IR models. Please list two methods to model term dependencies. Any two methods are welcome. (10 points)

3. Entity retrieval is a very common information need in IR. Term project one aims to deal with this issue. Please address four possible categories of queries to retrieve entities. (10 points)

4. Authors (i.e., writers of documents) and searchers (i.e., users of IR systems) have their own models to formulate documents and queries, respectively. Please discuss how to use these two models to rank documents for an input query. Moreover, please show how feedback documents can refine the retrieval results. (10 points)

5. Given a query q and two ranking models m_1 and m_2, assume m_1 and m_2 produce the following two ranking lists, respectively:

   A_{m1}=d39, d10, d3, d20, d44, d5, d30, d40, d50, d60

   A_{m2}=d70, d9, d56, d71, d25, d3, d89, d123, d80, d5

   Assume 10 documents shown as follows are judged and 4 of them are relevant (R) and 6 of them are non-relevant (NR).

   R={d3, d5, d9, d25}

   NR={d39, d44, d56, d71, d89, d123}

   (a) Please use the above example to describe the issue of the incomplete information in IR evaluation. (4 points)

   (b) Which model performs the best under the metric BREF? Please show their BREFs. (6 points)

6. In cross-lingual information retrieval (CLIR), a query in a language may be used to access documents in another language. An intuitive method for CLIR is to translate the input query to document language first and then perform the retrieval. Please propose a method which can do cross-lingual information retrieval without query translation. (10 points)

7. To prepare a labelled dataset is necessary for some researches. In IR, to label a document as relevant or non-relevant related to a query may need human annotators. Assume we hire three annotators to mark the relevancy of each candidate document. The majority of the three annotators determine the label of the document. Finally, a ground truth dataset is formed. Please propose a method to determine the performance of each annotator. (10 points)

8. Query terms in a query not occurring in a document will result in zero probabilities in IR modeling. Please discuss how statistical language modeling and neural network language modeling deal with this problem. (10 points)

9. In latent semantic indexing, we try to map both terms and documents into lower dimensional space, and perform the similarity computation on that space. Please show how to compute term vectors, document vectors, and vectors of input queries in the reduced space. (10 points)

10. Representations of queries and documents and computation of their relationship degree are key components in an IR framework. Please compare counting-based IR model and prediction-based IR model from aspects of representations and similarity computation. You can select any one model to describe your answers. (10 points)

11. Searching and routing are two sides of a same coin. In the class, you have learned lessons to evaluate the performance of a searching system. Please describe the concept of routing first, and discuss how to evaluate the performance of a routing system. (10 points)

12. Plagiarism (抄襲) can be defined as use or close imitation of the language and thoughts of another author and the representation of them as one's own original work (https://en.wikipedia.org/wiki/Plagiarism). Given a document and a reference document corpus, a plagiarism detection (抄襲偵測) task is defined to tell if the given document has some passages quite similar to some parts in the reference document set. Please propose a method to deal with this problem. Your method should identify which parts have problems and their sources. (10 points)