資格考試科目：資訊檢索與擷取

1. (26 pts) It is important to manually collect relevance assessments to evaluate the performance of information retrieval (IR) systems. But such process is usually expensive and time-consuming. One common solution is the pooling method. Assume that two human judges used the pooling method to evaluate the performance of ten IR systems. The following table shows how they rated the relevance of a collected pool of 20 documents to a certain query topic, in which R indicates relevance and N indicates non-relevance. Suppose that a document is considered relevant only if the two judges agree together in the evaluation.

<table>
<thead>
<tr>
<th>Doc ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>Judge 2</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Here are the top 10 ranking lists returned by two of the ten systems, NTU-1 and NTU-2, respectively, for this topic. Please answer the following questions.

<table>
<thead>
<tr>
<th>NTU-1</th>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc ID</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>16</td>
<td>17</td>
<td>12</td>
<td>19</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NTU-2</th>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc ID</td>
<td>4</td>
<td>17</td>
<td>3</td>
<td>11</td>
<td>13</td>
<td>16</td>
<td>17</td>
<td>15</td>
<td>6</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

(a) (6 pts) Please specify the pooling method and give its disadvantages.
(b) (4 pts) Explain why the pooling method is effective in performance evaluation.
(c) (4 pts) Calculate the kappa measure between the two judges.
(d) (6 pts) Calculate MAP (Mean Average Precision) and F1 of the NTU-1 system.
(e) (6 pts) Calculate 11-point interpolated precision (i.e., recall level precision) and 5-point document level precision (at 2, 4, 6, 8, and 10 documents) for the NTU-2 system.

2. (24 pts) Language modeling is vital for many research tasks such as speech recognition and machine translation. Language model (LM) is also well applied to IR. Please answer the following questions.

(a) (4 pts) Please prove that in the query likelihood model, ranking by \( p(d|q) \) is equivalent to ranking by \( p(q|d) \), where \( p(d|q) \) means that the probability of document \( d \) is interpreted as the likelihood that \( d \) is relevant to query \( q \), and \( p(q|d) \) calculates the probability of query \( q \) under the LM derived from document \( d \). Explain each step of your proof.
(b) (4 pts) Please give the physical meaning of the following query likelihood model and explain why it needs \( p(w_i | C) \).
\[ p(q \mid d) = \prod_{w \in q} \left[ \lambda \cdot p(w_i \mid d) + (1 - \lambda) \cdot p(w_i \mid C) \right], \]
where \( q \) and \( d \) stand for query and document, respectively, \( w_i \) denotes a word in the vocabulary, \( C \) is the document collection and \( \lambda \) is a weight varied from 0 to 1.

(c) (8 pts) Describe how to apply relevance feedback to LM.
(d) (4 pts) Describe the difference between LM and the probabilistic model.
(e) (4 pts) Please justify the following statement. Explain your answer.
"The vector space model is more efficient than the language model in speed."

3. (16 pts) Term mismatching has a great impact on the recall of most IR systems. Query expansion and relevance feedback are two of the most used approaches to deal with this problem.
(a) (4 pts) What is the difference between relevance feedback and query expansion?
(b) (4 pts) Why is positive feedback likely to be more useful than negative feedback to an IR system?
(c) (8 pts) Click-through data is the information collected by search engines in query logs. Such data records which hyperlinks in search results returned from search engines are clicked by users for certain queries. Please present a method that uses the click-through data to perform query expansion. Your method should deal with the sparseness problem that some queries entered by users do not appear in the logs. Explain why your method works.

4. (20 pts) The task of document classification is to assign a document to one or more categories based on its content. A general classification system often selects a subset of effective terms in training data as features and develops algorithms like NB (naive bayes), kNN (k Nearest Neighbor), SVM (Support Vector Machine) and Rocchio to determine which class(es) a given document belongs to.
(a) (6 pts) Please give two feature selection methods.
(b) (6 pts) Please explain why kNN handles multimodal classes better than Rocchio.
   The multimodal classes are those classes consisted of more than one clusters.
(c) (8 pts) Please show that multinomial NB model is formally identical to multinomial unigram LM.

5. (14 pts) Most blog websites allow users to assign tags to their blog posts. Such tags provide an easy way to categorize and search what they describe. However, many blogs still have no tags.
(a) (7 pts) Please propose a method that automatically extracts appropriate keywords as tags from textual content of blog posts. Explain why your method might work.
(b) (7 pts) Please propose a method that automatically generates good tags, where the tags cannot be extracted from textual content of blog posts. You can make use of other resources like the Web. Explain why your method might work.