Sample exam questions
COMP90042 Web Search And Text Analysis

May 20, 2015

In preparing for the exam, please consult the library collection of past exams at:

http://library.unimelb.edu.au/examination_papers

Specifically you should look for the exams from:

1. COMP90042 Web Search And Text Analysis, 2014

Note that the subject content has changed substantially, and only some of the questions from the above papers are relevant.

For this reason I’ve provided you with some indicative questions. Note that these are only indicative of the style of questions, and that the exam may cover other areas of the subject not covered by these example questions. I’ll publish answers to these in the last week of term, but I suggest you attempt these on your own first.

A: Short answer  (worth 14/50)

Expect to answer in a sentence or two, with longer or more thought-out answers corresponding to higher mark allocations.

1. What are the characteristics of a “Markov chain”? Name two examples where “Markov chains” are used in a retrieval or text analysis setting.  [1.5 marks]
2. What is an “information need” and how does this relate to a “query”? Use an example to justify your answer.  [1 mark]
3. Outline an approach to compression used in information retrieval, and state its effect in terms of time and space complexity.  [2 marks]
4. Describe why recall is important in IR evaluation, and why it is difficult to measure.  [2 marks]
5. The “Okapi BM25” model of document retrieval includes two additional components besides term frequency and inverse document frequency. State what these components are, and describe their respective roles in document ranking (no formulae needed).  [2 marks]
6. Describe the motivation behind the “n-gram” method for “language modelling” and name a limitation of this technique.  [2 marks]
7. Tree-structured models are often used for describing syntax. Describe a property of language that motivates the use of tree based techniques.  [1 mark]
8. Define with the aid of an example the term “hyponym”.  [1 mark]
B: Method questions  

These tend to require longer answers, again see the mark allocations for approximate answer difficulty and length.

Phrasal IR

1. Give the basic steps involved in positional index construction and querying for document retrieval with phrasal queries. [2 marks]

2. What is the space and time complexity of positional index construction? Justify your answer. [2 marks]

3. Identify one disadvantage and one advantage that the “positional index” has over a “bi-word index”. [1 mark]

Link analysis

1. The “page rank” method is framed as the frequency with which a random surfer visits each web page in a collection. Show mathematically how this gives rise to the model formulation and the iterative update equations. [3 marks]

2. Show mathematically that the update equations for the “page rank” method give rise to an Eigenvalue problem. Include your working. [2 marks]

3. Outline a way in which “page rank” scores could be used as part of a ranked document retrieval system. [1 mark]

Markov models

1. Contrast the use of $n$-gram Markov models with hidden Markov models. What is it about hidden Markov models that is “hidden”? [2 marks]

2. Present the Viterbi algorithm for a first order Hidden Markov Model with the aid of a simple example sentence, and state its time complexity of inference. [4 marks]

Machine translation

1. Define “word alignment”, and explain why it is often a necessary step in learning a machine translation system. [2 marks]

2. Contrast word-based and phrase-based translation, and provide two reasons why the phrase-based approach is more effective. [2 marks]
C: Algorithmic Questions (worth 10/50)

Given the following term-document matrix

<table>
<thead>
<tr>
<th>doc</th>
<th>“soccer”</th>
<th>“football”</th>
<th>“pitch”</th>
<th>“hockey”</th>
<th>“tournament”</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

1. compute the cosine similarity between the query “soccer” and each document, using the vectors above (no need to include IDF term), and show the ranked order of documents. You are not required to simplify fractions. [2 marks]

2. using the Rocchio algorithm, defined as

\[ q_e = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_i \in D_r} d_i - \gamma \frac{1}{|D_{nr}|} \sum_{d_i \in D_{nr}} d_i \]

compute the new query vector for “soccer” using the top ranked document for pseudo relevance feedback (with \( \alpha = \beta = 0.5, \gamma = 0 \)) and compute the new document ranking. You are not required to simplify fractions. [3 marks]

D: Essay (worth 8/50)

You’ll be required to write about a page on one of three options. This is a chance to display your broader understanding of the taught material and how it fits together. Here’s a sample topic:

**Query Expansion** In retrieval, query expansion is the approach of automatically adapting a user’s query to produce different retrieval results. Provide the motivations for doing query expansion, compare several methods for doing query expansion, and outline their strengths and weaknesses. Consider in your comparison the baseline method using the query directly, without expansion. [8 marks]

**Concluding remarks**

Note that the above questions don’t cover every topic in the subject, so please don’t read too much into the areas covered. You will need to prepare on the full range of topics covered in the lectures and workshops. Note also that the exam will be significantly longer than this, particularly in parts B and C. You can get a ballpark estimate by looking at the number of marks assigned to each section (worth XX/50).