The University of Melbourne

Department of Computing and Information Systems

COMP90042
Web Search and Text Analysis
June 2015

Identical examination papers: None

Exam duration: Two hours

Reading time: Fifteen minutes

Length: This paper has 5 pages including this cover page.

Authorised materials: None

Calculators: Not permitted

Instructions to invigilators: Students may not remove any part of the examination paper from the examination room. Students should be supplied with the exam paper and a script book, and with additional script books on request.

Instructions to students: This exam is worth a total of 50 marks and counts for 50% of your final grade. Please answer all questions in the script book provided, starting each question on a new page. Please write your student ID in the space below and also on the front of each script book you use. When you are finished, place the exam paper inside the front cover of the script book.

Library: This paper is to be held in the Baillieu Library.

Student id:

Examiner's use only:

Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*

COMP90042 Web Search and Text Analysis
Final Exam

Semester 1, 2015

Total marks: 50

Students must attempt all questions

Section A: Short Answer Questions [14 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in no more than a line or two.

Question 1: General Concepts [6 marks]

1. The frequency of words in natural language is often described as having a "Zipfian distribution". Describe what this means, and outline an important consequence for text processing. [2 marks]
2. Provide a definition of the "mean average precision" evaluation measure, and justify why it is appropriate for evaluating ranked retrieval. [2 marks]
3. "Smoothing" is often important when dealing with models of text; outline two contexts where "smoothing" is used and describe its effect. [2 marks]

Question 2: Information Retrieval [4 marks]

1. What are "stop-words" and why are they often discarded in information retrieval? [2 marks]
2. Explain the reason why "posting lists" are typically stored in sorted order by document identifier. [1 mark]
3. Compression is often considered in the context of efficient on-disk storage. Describe why compression is also important for in-memory storage, in the context of information retrieval. [1 mark]

Question 3: Text Analysis [4 marks]

1. Natural language is ambiguous in many ways; provide two examples illustrating different kinds of ambiguity in English. [2 marks]
2. What is a "homonym" and why might they prove problematic for language processing? [1 mark]
3. Outline with the aid of an example how the "IOB" method allows for sequence classification to be applied to multi-word labelling tasks such as named entity recognition. [1 mark]

Section B: Method Questions [18 marks]

In this section you are asked to demonstrate your conceptual understanding of the methods that we have studied in this subject.

Question 4: Document Search [7 marks]

1. Present and contrast the algorithms for the "bi-word" and "positional" index methods for document retrieval with phrasal queries. For both methods show the algorithms for querying and index construction, and illustrate their operation with a simple example. [4 marks]

2. Outline a method for compressing a positional index, and describe what property of language is being exploited by this technique. [3 marks]

Question 5: Web as a Graph [4 marks]

The "page rank" and "hubs and authorities" algorithms are means of deriving document importance measures automatically from a hyper-linked corpus.

1. Explain with the aid of an example the terms "hubs" and "authorities". [1 mark]

2. The "page rank" method is framed as the frequency with which a random surfer visits each web page in a collection. Provide a similar analogy for the "hubs and authorities" method, and show how this gives rise to a probabilistic model with iterative update equations,

   $$h_i \leftarrow \sum_{i \to j} a_j$$
   $$a_i \leftarrow \sum_{j \to i} h_j$$

   where i and j index the pages, $i \to j$ and $j \to i$ denote directed edges in the graph (hyperlinks) and $\mathbf{h}$ and $\mathbf{a}$ are vectors of hub and authority scores. [3 marks]

Question 6: Markov Models [7 marks]

1. Describe the assumptions that underlie Markov models, and provide a part-of-speech tagging example showing where these assumptions are inappropriate. [2 marks]

2. What classes of formal languages can be described by Markov models over word sequences? Relate this to context free grammars used in parsing. [2 marks]

3. The Viterbi algorithm for hidden Markov models uses dynamic programming to compute the maximum probability path of hidden states that generates a given sequence of observations,

   $$\mathbf{t}^* = \text{argmax}_{t_1,t_2,\ldots,t_{N-1},t_N} \, p(w_1, t_1, w_2, t_2, \ldots, w_{N-1}, t_{N-1}, w_N, t_N)\,.$$

   Consider the related *forward* problem of marginalising (summing) the probability over all paths for an observation sequence in order to compute the probability of the observations, i.e.,

   $$p(w_1, w_2, \ldots, w_{N-1}, w_N) = \sum_{t_1,t_2,\ldots,t_{N-1},t_N} p(w_1, t_1, w_2, t_2, \ldots, w_{N-1}, t_{N-1}, w_N, t_N)\,. \qquad (1)$$

   Using a similar approach to the Viterbi algorithm, show how Equation (1) can also be solved using dynamic programming. This involves reducing Equation (1) to a recursive formulation. [3 marks]

Section C: Algorithmic Questions [10 marks]

In this section you are asked to demonstrate your understanding of the methods that we have studied in this subject, in being able to perform algorithmic calculations.

Question 7: Document and Term ranking [5 marks]

Consider the following "term-document matrix", where each cell shows the frequency of a given term in a document:

DocId	stock	share	corporate	corruption	london	barclays
doc_1	2	2	1	0	0	0
doc_2	1	3	2	0	1	1
doc_3	0	0	1	1	1	1
doc_4	0	0	0	3	4	0

This question is about using the term-document matrix for querying and to find similar terms.

1. Calculate the "document ranking" for the query *corporate corruption* using a "cosine similarity" measure over raw term frequencies. [1.5 marks]

2. Calculate the "retreival status value" for doc_2 with the query *corporate corruption* according to a smoothing unigram language model, $P(d|q) \propto \prod_{t \in q} \lambda P(t|M_d) + (1 - \lambda) P(t|M_c)$. Use maximum likelihood estimates for the document specific language models M_d and corpus language model M_c and let $\lambda = 0.5$. You do not need to simplify numerical values. [2 marks]

3. Calculate the pairwise similarity between term *corporation* and all other terms (*stock*, *share*, ..., *barclays*) based on "cosine similarity". You do not need to simplify numerical values. [1.5 marks]

Question 8: Grammars and Parsing [5 marks]

Describe an algorithm for chart parsing with a context free grammar. Illustrate with a worked example, e.g., using a simple grammar and a short 4-7 word sentence.

Section D: Essay Question [8 marks]

Question 9: Essay [8 marks]

Discuss *one* of the following options (about 1 page). Marks will be given for correctness, completeness and clarity.

- **Word sense ambiguity.** Define the problem of word sense ambiguity, with the aid of examples. Motivate why this is an important problem and a hard one to solve, and outline methods for word sense disambiguation.
- **Machine Translation.** A long running challenge in language processing has been the automatic translation between different languages. Discuss the key difficulties of translation, outline the sub problems and how they can be solved to create an automatic translation system, and discuss the strengths and weaknesses of these solutions.
- **Relevance feedback.** User interactions are a key source of feedback in information retrieval. Describe both pseudo- and regular relevance feedback, and outline ways in which relevance feedback can be used in a retrieval system. Describe the strengths and weaknesses of relevance feedback compared with standard one-shot retrieval and supervision through click-through data and query log mining.

— End of Exam —

Author/s:

Computing and Information Systems

Title:

Web search and text analysis, 2015 Semester 1, COMP90042

Date:

2015

Persistent Link:

http://hdl.handle.net/11343/90921