COMP90042 Web Search and Text Analysis
Sample Exam

Semester 1, 2016

Total marks: 50 (roughly)

Students must attempt all questions

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, 2015
2. 433-{4,6}60 Human Language Technology, 2003, 2004, 2008.

Note that the subject content has changed substantially from prior years, and only some of the questions from the above papers are relevant. Last year's exam is the most relevant, and the format will be the same.
I've provided you with some further indicative questions below.

A: Short answer

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

1. Information retrieval

   a) What is an "information need" and how does this relate to a "query"? Use an example to justify your answer. [2 marks]

   b) Outline an approach to compression used in information retrieval, and state its effect in terms of runtime. [2 marks]

   c) Describe why recall is important in IR evaluation, and why it is difficult to measure. [2 marks]

   d) 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]

2. Probabilistic models of language

   (a) What are the characteristics of a "Markov chain"? Name an example where "Markov chains" are used in a retrieval or text analysis setting. [2 marks]

   (b) Describe the motivation behind the "n-gram" method for "language modelling" and name a limitation of this technique. [2 marks]

   (c) Tree-structured models are often used for describing syntax. Describe a property of language that motivates the use of tree based techniques, and provide a supporting example. [2 marks]

   (d) Define with the aid of an example the term "hypernym". [1 mark]

   (e) Distributional Semantics

      i. Give one reason why "PMI" would be preferred over "term frequency" as a measure of word association for computational semantics. [1 mark]

ii. The creation of “word2vec” word vectors (embeddings) applies supervised classification techniques. What is the underlying classification task for each of the two approaches we discussed (that is, “CBOW” and “skip-gram”)? [1 mark]

iii. What is “dimensionality reduction”? Give two examples [2 marks]

B: Method questions

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

1. Markov models

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

   b) 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. [5 marks]

2. Machine translation

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

   b) Contrast word-based and phrase-based translation, and provide a reason why the phrase-based approach is more effective. [2 marks]

3. Parsing

   (a) What is the difference between the “CYK” and “Earley” parsing algorithms with respect to the expectations of the form of the input context free grammar. You should give the explicit restrictions. [2 marks]

   (b) Give a slightly modified version of the more-restricted algorithm which does not have these restrictions (this will involve an extreme increase in the time complexity of the algorithm). In doing this, you should explain the key part of the original algorithm [2 marks]

   (c) Explain how the other algorithm (the one without restrictions) avoids the blow up in complexity seen in the proposed modification above. [1 mark]

C: Algorithmic Questions

Rocchio algorithm Given the following term-document matrix

	Term frequency, $w_{t,d}$				
doc	“soccer”	“football”	“pitch”	“hockey”	“tournament”
d1	3	0	4	0	0
d2	0	6	8	0	0
d3	1	0	0	2	2

a) 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. [3 marks]

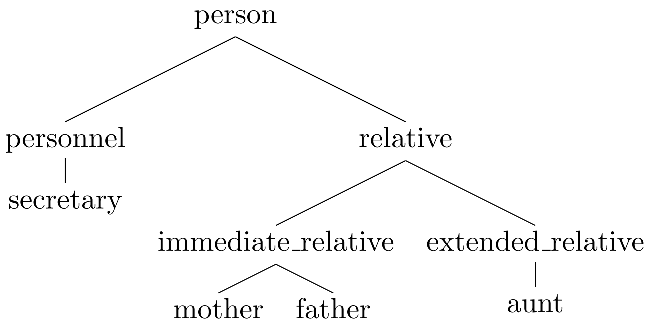
Continued overleaf ...

b) 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]

Lexical Semantics This question is based on the following lexical hierarchy.



1. Rank the all other terms with respect to their path similarity to *relative*, and mention one result that doesn't seem reasonable. [1 mark]
2. Calculate Wu-Palmer similarity between all the leaves of the tree. You can leave the results in fraction form. [2 marks]
3. In a corpus, we find that 1 out of every 16 words is a *person*, 1 out of every 128 words is a *relative*, 1 out of every 512 words is a *relative.* 1 out of every 1024 is a *mother*, and 1 out of every 512 is a *personnel.* What's the Lin distance between *mother* and *personnel* based the corpus statistics and the given hierarchy? [2 marks]

D: Essay

You'll be required to write about a page on one of four options. See last year's exam for some ideas.

Concluding remarks

Note that the above questions don't cover every topic in the subject (some whole areas are missing!), 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 question versus the section totals on last years' exam (worth XX/50).

— *End of Exam* —