The University of Melbourne

Department of Computing and Information Systems

COMP90042
Web Search and Text Analysis
June 2018

Identical examination papers: None

Exam duration: Two hours

Reading time: Fifteen minutes

Length: This paper has 6 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*

Continued overleaf . . .

COMP90042 Web Search and Text Analysis

Semester 1, 2018

Total marks: 50

Students must attempt all questions

Section A: Short Answer Questions [15 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 [8 marks]

a) Explain why the F_1 score is widely used in evaluating text processing methods, but is much less widely used in retrieval. [1 mark]

b) Name one task where the "bag-of-words" document representation is appropriate, and another task where "bag-of-words" is inappropriate. In each task, explain why this is the case. [2 marks]

c) Contrast dependency parsing with syntactic phrase-structure parsing. Identify at least one key similarity and one important difference. [2 marks]

d) Explain why "word alignment" in machine translation is typically framed as an unsupervised learning problem. [1 mark]

e) Unsupervised HMMs can be trained using Expectation-Maximisation. Explain the difference between hard EM and soft EM. Why is soft EM considered better? [2 marks]

Question 2: Information Retrieval [4 marks]

a) Name a situation where "compression" is important in information retrieval, and for your chosen situation, explain what property of the data allows for compression. [1 mark]

b) What is pseudo relevance feedback and why is it useful? [1 mark]

c) The "mean average precision" metric is often used for evaluation of retrieval effectiveness. State its formulation, and explain why it is more appropriate than precision at rank k. [2 marks]

Question 3: Distributional Semantics [3 marks]

a) Explain two important benefits of using vector representations for words, instead of simple discrete features. [2 marks]

b) The "Word2Vec skip-gram" algorithm and "latent semantic analysis" both learn vector representations of words. Explain a key difference between these two approaches. [1 mark]

Continued overleaf . . .

Section B: Method Questions [17 marks]

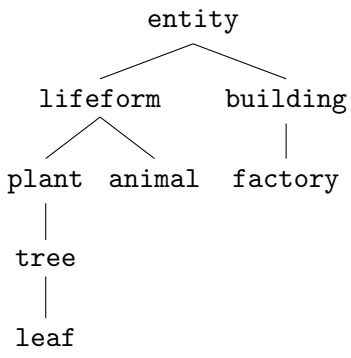
In this section you are asked to demonstrate your conceptual understanding of the methods that we have studied in this subject.

Question 4: Lexical semantics [5 marks]

a) Fill in this sentence with the appropriate *-nym*: *leaf* is a _____ of *plant* (in the sense of vegetation). [1 mark]

b) Given the hierarchy below calculate the following similarity metrics: [2 marks]

- Path similarity:
  - `plant` and `factory`
  - `leaf` and `animal`
- Wu-Palmer similarity:
  - `plant` and `factory`
  - `leaf` and `animal`



c) How is "Lin similarity" calculated? Mention all resources required and the major steps in the calculation. [2 marks]

Question 5: Grammars and Parsing [7 marks]

This question is based on the "CYK" algorithm for PCFGs, shown below in pseudo-code:

```
function parse-CYK(w, G):
1        for j in 1 .. |w|
2            for all A -> wj in grammar
3                set chart[j-1,j,A] = P(A -> wj)
4            for i in j-1 .. 0 (descending)
5                for k in i+1 .. j-1
6                    for all A -> B C in grammar
7                        prob = P(A -> B C) * chart[i,k,B] * chart[k,j,C]
8                        if prob > chart[i,j,A] then
9                           chart[i,j,A] = prob
10                          back[i,j,a] = (k, B, C)
11       return build-tree(back, |w|, S)
```

a) The "CYK algorithm" assumes the grammar is in "Chomsky Normal Form (CNF)". What is CNF? Explain with relation to the above code, why this restriction is required. [2 marks]

Continued overleaf . . .

b) What is stored in the `chart`? [1 mark]

c) Explain how the iteration order in the for-loops above is critical for the correct updating of the chart. [2 marks]

d) Provide code for the algorithm `build-tree`, which recovers the best scoring tree, as invoked on line 11. Feel free to use Python or pseudo-code syntax. [2 marks]

Question 6: Information Extraction [5 marks]

Consider the following toy data, composed of only one sentence, its corresponding Named Entity annotation and a gold set of relations extracted from it:

- `Amanda Palmer is a singer-songwriter born in 1976 in New York City, New York, US.`
- `[Amanda Palmer]`$_{\mathbf{PER}}$ `is a singer-songwriter born in [1976]`$_{\mathbf{TIME}}$ `in [New York City]`$_{\mathbf{LOC}}$`, [New York]`$_{\mathbf{LOC}}$`, [US]`$_{\mathbf{LOC}}$`.`
- Gold relations:
  - `year-of-birth(Amanda Palmer, 1976)`
  - `place-of-birth(Amanda Palmer, New York City)`
  - `city-state(New York City, New York)`
  - `state-country(New York, United States of America)`

a) Suppose you want to train a Named Entity Recogniser using an HMM. Rewrite the NE annotated sentence into a sequence of (word, tag) elements using one of the schemes you learned in class. Write your answer in the following format: `word1/tag1 word2/tag2 ...` [2 marks]

b) The first step in Relation Extraction is to build a binary classifier that recognises if two entities have a relation or not. Assuming the example above is the only data you have available, how many positive and how many negative instances you would have in your training set for this classifier? [1 mark]

c) The second step in Relation Extraction is to build a multi-class classifier that, given a positive entity pair, predicts the relation between them. However, even if you have a perfect classifier the relations extracted from the sentence will not match the gold relations given above. Why is this the case? How would you solve this problem, so the relations match the gold standard? [2 marks]

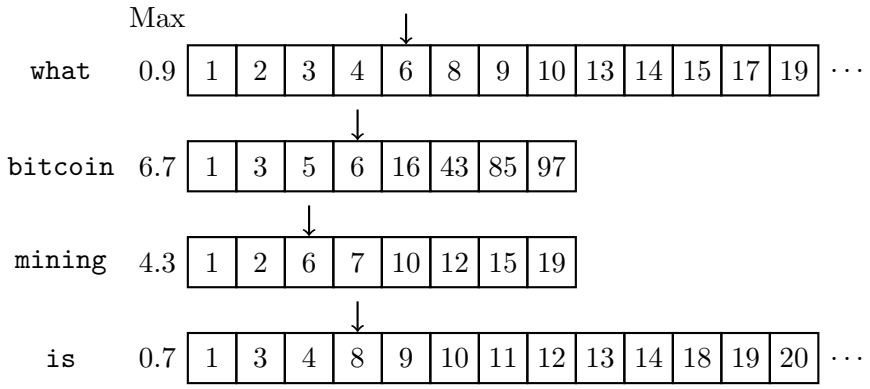
Continued overleaf ...

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 [4 marks]

Consider the following snapshot of an instance of the WAND top-K query processing algorithm for the query `'What is bitcoin mining'` and $K = 2$. Suppose the algorithm has just evaluated document 6 with score 4.4. Answer the following questions:



#	Score	DocId
1	8.7	1
2	4.1	3

a) What is the idea that allows the WAND algorithm to skip evaluating documents? [1 mark]

b) Was document 4 ever in the top-K score list? Explain your reasoning. [1 mark]

c) What is the next document that will be evaluated? Explain your reasoning. [2 marks]

Question 8: N-gram language modelling [6 marks]

This question asks you to calculate the probabilities for n-gram language models. You should leave your answers as fractions. Consider the following corpus with 3 "sentences":

bbaccab
aaacbba
bacabba

a) Calculate a unigram language model for this corpus. [1 mark]

b) Next, calculate an MLE bigram language model. Add additional symbols as needed. Under this model, what is the probability of the sentence *abc*? [3 marks]

c) Calculate smoothed bigram probabilities for all those terms where b is the context (that is, $p(\cdot|b)$) by applying add-one smoothing. Calculate the probability of the sentence *abc* under this new model. [1 mark]

d) Calculate smoothed bigram probabilities for terms where b is the context (that is, $p(\cdot|b)$) by interpolating it with unigram counts, using 0.5 as the interpolation weight. Calculate the probability of the sentence *abc* under this new model. [1 mark]

Continued overleaf ...

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.
- **Word vector learning.** Define the problem of using discrete word representations, with the aid of examples. Motivate why this is an important problem, outline methods used to solve it, how it is evaluated and how word vectors learning, and word vectors themselves, are related to other tasks in language processing.
- **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, how they are evaluated, and discuss the strengths and weaknesses of these solutions.

— End of Exam —