The University of Melbourne

Department of Computing and Information Systems

COMP90042

Web Search and Text Analysis

June 2017

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 (but not sub-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*	*Q10*

COMP90042 Web Search and Text Analysis
Final Exam

Semester 1, 2017

Total marks: 50

Students must attempt all questions

Section A: Short Answer Questions [10 marks]

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

Question 1: General Concepts [4 marks]

a) What is a key difference between "lemmatisation" and "stemming"? [1 mark]

b) State the goal that guides the training of a"logistic regression" classifier (as well as related models like Word2Vec). [1 mark]

c) Name one advantage associated with "dimensionality reduction" for representing text. [1 mark]

d) Describe the "IOB" tagging method used in information extraction, and explain why it is a useful technique. [1 mark]

Question 2: Information Retrieval [3 marks]

a) What is the intuition behind the use of "tf" and "idf" factors in ranked retrieval? [1 mark]

b) Give an example of a query that will result in a poor (very long) runtime of the standard vector-space model querying algorithm, and explain why this is the case. [1 mark]

c) The *page rank* and *hubs and authorities* (HITS) methods exploit the link structure of the web. What is the common intuition behind these methods? [1 mark]

Question 3: Discourse [3 marks]

a) How is the problem of "discourse segmentation" usually framed to convert it into a standard classification task? [1 mark]

b) Give an example of a "discourse marker" and an "RST discourse relation" that it indicates. [1 mark]

c) State one property which always holds between an "anaphor" and its "antecedent", and another which often holds. [1 mark].

Section B: Method Questions [14 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) What is the term which refers to two unrelated senses of the same word type? [1 mark]

b) Build a WordNet-style graph of lexical relations using the following lemmas: *plant*, *entity*, *leaf*, *lifeform*, *building*, *tree*, *factory*, *animal*. Circle lemmas to indicate synsets. Hypernym and meronym relations should be distinguished. [2 marks]

c) Propose an unsupervised algorithm for word sense disambiguation which makes use of graphs of lexical relations. Use the graph you just created to provide an example of how it would work (you can add nodes if needed). [2 marks]

Question 5: Dependency parsing [6 marks]

a) Describe what it means for two words to be in a "dependency" relation, and provide an example. [1 mark]

b) Show the dependency parse for the sentence

> Donald always speaks the truth.

You do not need to provide edge labels. [1 mark]

c) Show a sequence of parsing steps using a "transition-based parser" that will produce this dependency parse. [2 marks]

d) Some dependency trees are "non-projective". Explain what this means, and then explain why the "transition-based parser" algorithm can only create projective trees. [2 marks]

Question 6: Markov Models [3 marks]

a) State the probabilistic formulation of an "n-gram language model", and use this to explain why they are a type of "Markov model". [1 mark]

b) Show how a bigram language model can be implemented as a *probabilistic context free grammar (PCFG)*. State what productions are needed, and their corresponding weights. [1 mark]

c) The "page rank" algorithm is a form of Markov chain. State the probabilistic formulation of page rank, and state how the parameters are defined using the hyperlink structure. [1 mark]

Section C: Algorithmic Questions [18 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 Search [7 marks]

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

DocId	alternative	fact	lie	truth	trump
doc$_1$	1	1	0	0	2
doc$_2$	0	2	0	3	0
doc$_3$	1	1	3	1	3
doc$_4$	0	1	0	1	1

We will be using the following query, alternative fact lie, and the vector-space model of retrieval.

a) Illustrate the "inverted index" for this document collection, showing both the "postings lists" and "document frequencies". In this part, use raw term frequencies. [1 mark]

b) Compute the "IDF" term for each of the query terms. Use the standard logarithmic formulation of IDF, with base 2 logarithms (logarithm table provided below). [1 mark]

c) Using your answers to the above, illustrate the progress of the querying algorithm in the TF*IDF vector space model for the given query. Show the state of the accumulator after each step, and the final document ranking. To keep things simple, do not perform vector length normalisation. [1 mark]

d) Assuming the final ranking was

   doc$_4$, doc$_2$, doc$_3$, doc$_1$

   and we have the following manual "relevance judgements (qrels)"

   doc$_1$: 1, doc$_2$: 1, doc$_3$: 0, doc$_4$: 0

   where 1 means the document relevant is relevant to the query, and 0 means irrelevant. Compute the "average precision" and "F_1 score". Show your working. [2 marks]

e) Show the results of "posting list compression", with the "variable byte compression" and "delta" method applied to the posting list:

   5, 69, 70, 326, 329, 16714 .

   Report the resulting byte sequence, using the integer value for each byte (e.g., 65 corresponds to 01000001). [2 marks]

The following values may be useful:

x	0	1	2	3	4	5	6	7	8	9
$\log_2 x$	-	0.0	1.0	1.6	2.0	2.3	2.6	2.8	3.0	3.2
2^x	1	2	4	8	16	32	64	128	256	512
x	10	11	12	13	14	15	16	17	18	19
$\log_2 x$	3.3	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2	4.2
2^x	1024	2048	4096	8192	16384	32768	65536	131072	262144	524288

Question 8: Machine Translation [5 marks]

This question is about word based models of machine translation. Consider the following English-Spanish parallel corpus, comprising three sentence pairs:

purple haze	grey door	grey haze
neblina púrpura	puerta gris	neblina gris

You will be showing how IBM model 1 is trained from parallel texts. This model defines the probability of a sentence translation as

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(1+I)^J} \prod_{j=1}^{J} t(f_j|e_{a_j})$$

where I and J denote the lengths of the English and foreign sentences, respectively, i.e., $I = |\mathbf{e}|$ and $J = |\mathbf{f}| = |\mathbf{a}|$, and $t(f|e)$ are the parameters of the model which describe the translation from an English word e to a foreign word f.

a) Starting with the parameters, $t(f|e)$ set to uniform probabilities, show the "expected counts" over word alignments the training corpus (the 'E' step in EM). You should exclude NULL alignments. [2 marks]

b) Now use your expected counts to derive updated values for the parameters $t(f|e)$ (the 'M' step in EM). For this question, just show the conditional probabilities for translating $e = grey$ and $e = door$ into Spanish. [2 marks]

c) Repeating the above E and M steps will lead the EM algorithm to learn the correct alignments for this data (puerta = door, gris = grey, etc.) Explain what property of the data is being exploited to arrive at this solution, despite starting with uniform parameters? [1 mark]

Question 9: Grammars and Parsing [6 marks]

This question is about using analyzing syntax. Consider the following ambiguous sentence:

```
Hit the man with the ball.
```

a) Describe the syntactic ambiguity in this sentence [1 mark]

b) Write a set of linguistically-plausible CFG productions that can represent and structurally differentiate the two interpretations. [2 marks]

c) Do an Earley chart parse of the sentence using your grammar. You should include the full chart, which will include the edges for both possible interpretations. [3 marks]

Section D: Essay Question [8 marks]

Question 10: Essay [8 marks]

Choose one of the three topics below, and discuss it in detail. At a minimum, your essay should do the following:

a) Define the topic, and motivate why it is important.

b) Explain how it relates to various tasks discussed in class. You should aim to cover 2-3 different tasks.

c) Discuss how its application is similar and differs across the tasks, and analyze why this is the case.

Marks will be given for correctness, completeness and clarity. Expect to write about 1 page.

- **Smoothing** (also known as **Regularisation**).
- **Evaluation Methods and Datasets.**
- **Vector Space Models.**

— End of Exam —