The University of Melbourne

Department of Computing and Information Systems

# COMP90042

# Web Search and Text Analysis

# June 2016

**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
# Final Exam

**Semester 1, 2016**

**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 two or three sentences.

### Question 1: General Concepts [5 marks]

a) How is logistic regression "regularisation" related to the concept of overfitting in supervised classification? [1 mark]

b) For higher order ($n \geq 2$) $n$-gram language models, what is the key idea that differentiates more sophisticated "smoothing" techniques from stand-alone add-$k$ smoothing? Mention one smoothing technique which instantiates this idea. [2 marks]

c) The "IBM model 1" of word-based translation is formulated as $P(F|E) = \sum_A P(F, A|E)$ where $E$ and $F$ are a parallel sentence pair and $A$ is the unknown word alignment. This model is trained using the "expectation maximisation" algorithm. State the two key steps in the training loop, and, for each step, the mathematical quantity being computed. [2 marks]

### Question 2: Information Retrieval [6 marks]

a) What is an "information need" and how does this relate to a "query"? Use an example to justify your answer. [1 mark]

b) What is the "bag-of-words" assumption, and why is it often used in information retrieval? [1 mark]

c) Describe the variable-byte compression method, and explain why this is useful for compact storage of an inverted index. [2 marks]

d) Provide two reasons why the "document frequency" is often stored explicitly for each term in an inverted index. [2 marks]

### Question 3: Distributional Semantics [3 marks]

a) Compare "Latent Semantic Analysis" (that is, LSA or term-document SVD) and "Latent Dirichlet Allocation" (LDA), identifying two important commonalities and two important differences. [2 marks]

b) The (Word2Vec) "skip-gram" model can be viewed as kind of matrix factorisation (or decomposition). Identify and explain one property of the matrix it (implicitly) factorises which differentiates it from LSA. [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: Text Classification [6 marks]

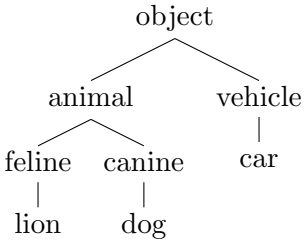
For this question, suppose you have a very large corpus of English texts written by people from 20+ different language backgrounds, and you want to build an automatic Native Language Identification system.

a) Name two types of "features" you think would be appropriate for this task and explain why. [2 marks]

b) Given the nature of the task and the features you have chosen, would you perform "lemmatisation" and/or "stop word removal" over your corpus? Explain why or why not for both preprocessing methods. [2 marks]

c) Given the task and the features you have chosen, do you think a Random Forest classifier would be appropriate? What about a Support Vector Machine? Justify your answers. [2 marks]

### Question 5: Lexical semantics [5 marks]




The questions below are based on the partial lexical hierarchy above.

a) Fill in this sentence with the appropriate *-nym*: animal is a _____ of lion. [1 mark]

b) Based on simple "path-based" similarity, which is more similar to lion, dog or vehicle? What about with the "Wu-Palmer" similarity metric? [2 marks]

c) If we are using "Lin" similarity, is it possible that lion might be more similar to car than it is to dog? If so, show give the condition on the "information content" of dog that must hold (in terms of the IC of other nodes) for this to happen, or, if not, explain why not. [2 marks]

### Question 6: Markov Models and Grammars [6 marks]

a) Hidden Markov models (HMMs) describe a probability distribution over both words $w_i$ and tags $t_i$. State the formulation of the joint probability, $p(w_1, w_2, \ldots, w_N, t_1, t_2, \ldots, t_N)$, for a first order hidden Markov model, and use your formulation to describe two of the modelling assumptions. [2 marks]

b) A first order hidden Markov model can be expressed with a probabilistic context-free grammar (PCFG), as follows:

| | | | |
|---|---|---|---|
| $S$ | $\rightarrow t\ X_t$ | [_____] | for all tags $t$ |
| $X_t$ | $\rightarrow t'\ X_{t'}$ | [_____] | for all tag pairs $t, t'$ |
| $X_t$ | $\rightarrow$ `</s>` | [_____] | for all tags $t$ |
| $t$ | $\rightarrow w$ | [_____] | for all tags $t$ and words $w$ |



Continued overleaf ...

where S is the start symbol, $X_t$ are a special non-terminal symbols (for each tag $t$), `</s>` is a special terminal denoting the the end of the sentence, and each rule has a probability denoted in square brackets, []. Fill in the empty boxes with the relevant parts of the probability formulation of the first order hidden Markov model (from your answer to part 1). [2 marks]

c) Using the above grammar illustrate the equivalence between the CYK algorithm for PCFG parsing and the Viterbi algorithm for the HMM. You might want to show an example sentence to aid your explanation. [2 marks]



Continued overleaf ...

## Section C: Algorithmic Questions [11 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 | snipe | tax | tony | boats | malcolm | panama |
|---|---|---|---|---|---|---|
| $\text{doc}_1$ | 2 | 1 | 0 | 0 | 1 | 1 |
| $\text{doc}_2$ | 0 | 0 | 3 | 2 | 1 | 0 |
| $\text{doc}_3$ | 2 | 0 | 0 | 0 | 1 | 0 |
| $\text{doc}_4$ | 0 | 3 | 4 | 0 | 2 | 0 |

a) Calculate the document ranking for the query *tax panama*, using the "TF*IDF" measure of similarity with the standard versions of TF (raw frequency) and IDF (logarithmic). Show your working. You do not need to simplify numerical values, and should use logarithms with base 2. [3 marks]

The following values may be useful:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sqrt{x}$ | 0 | 1.0 | 1.4 | 1.7 | 2.0 | 2.2 | 2.4 | 2.6 | 2.8 | 3.0 |
| $\log_2 x$ | - | 0.0 | 1.0 | 1.6 | 2.0 | 2.3 | 2.6 | 2.8 | 3.0 | 3.2 |
| $x$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| $\sqrt{x}$ | 3.2 | 3.3 | 3.5 | 3.6 | 3.7 | 3.9 | 4.0 | 4.1 | 4.2 | 4.4 |
| $\log_2 x$ | 3.3 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 | 4.1 | 4.2 | 4.2 |
| $x$ | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| $\sqrt{x}$ | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.0 | 5.1 | 5.2 | 5.3 | 5.4 |
| $\log_2 x$ | 4.3 | 4.4 | 4.5 | 4.5 | 4.6 | 4.6 | 4.7 | 4.8 | 4.8 | 4.9 |

b) Using query expansion with pseudo-relevance feedback, compute the new ranking. You should use the Rocchio algorithm with parameters $\alpha = \beta = 0.5$ and $\gamma = 1$, and treat the top ranked candidate as relevant. Show your working. [2 marks]

### Question 8: Part of speech, Grammars and Parsing [6 marks]

This question is about using analyzing syntax. Consider the following newspaper headline:

```
Eye drops off shelf
```

a) First show the key ambiguity in the sentence by giving two possible part-of-speech tag sequences. You can use any existing POS tagset, or your own, provided it satisfies the basic properties of a tag set and is easily interpretable. The tag set you use need not distinguish inflectional differences. [1 mark]

b) Write a set of CFG productions that can represent and structurally differentiate these two interpretations. Your set of non-terminals should consist of S, NP, VP, and your POS tag set from above, and your rules should have no recursion. [2 marks]

c) Do a CYK or Earley parse of the sentence using your grammar. You must include the full table/chart. If you choose to do CYK and have to convert your grammar to Chomsky Normal Form, be sure to show which productions must be changed. [3 marks]



Continued overleaf ...

# Section D: Essay Question [8 marks]

## Question 9: Essay [8 marks]

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

a) Define the task and identify any key subtasks or common variations.

b) Point out significant connections between the task and at least 2 other major topics discussed in this class. For full credit, at least one of these connections must be something we did not explicitly discuss in lecture.

c) Discuss how the task is evaluated. Mention any significant challenges to evaluation, standard test sets, and the metric(s) typically used.

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

- **Sentiment Analysis.**
- **Information Extraction.**
- **Question Answering.**
- **Machine Translation.**

*— End of Exam —*