The University of Melbourne

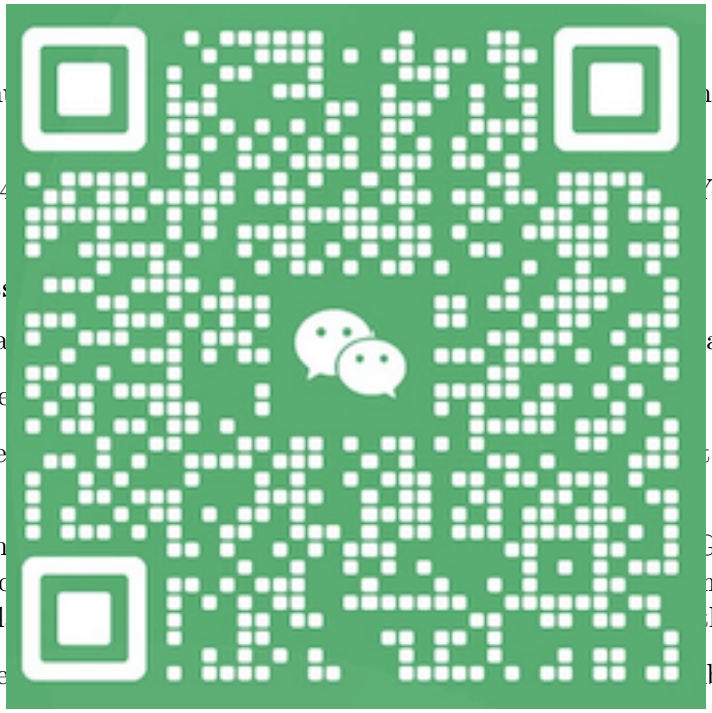
School of Computing and Information Systems

COMP90042

Natural Language Processing

Final Exam

Semester 1 2023



Exam duration: 165 min … e + 30 minutes upload time)

Length: This paper has 4 … You should attempt all questions.

Instructions to students

- This exam is worth a … ade.
- You can read the que…
- You are recommende… some answers require drawing diagrams or …
- You will need to scan … Gradescope. Be sure to label the scans/photo… nlabelled question, e.g. if 4 questions are unl… h a penalty of -10%).
- Please answer all que… ber on every page.

Format: Open Book

- While you are undertaking this assessment you are permitted to:
  - make use of the textbooks, lecture slides and workshop materials.
- While you are undertaking this assessment you must **not**:
  - make use of any messaging or communications technology;
  - make use of any world-wide web or internet-based resources such as Wikipedia, Stackoverflow, Google, AI services (e.g. ChatGPT) or other web services;
  - act in any manner that could be regarded as providing assistance to another student who is undertaking this assessment, or will in the future be undertaking this assessment.
- The work you submit must be based on your own knowledge and skills, without assistance from any other person.

Continued overleaf . . .

COMP90042 Natural Language Processing

Semester 1, 2023

Total marks: 120 (40% of subject)

Students must attempt all questions

Section A: Short Answer Questions [33 marks]

Answer each of the questions in this section briefly. Each answer should be no longer than several sentences.

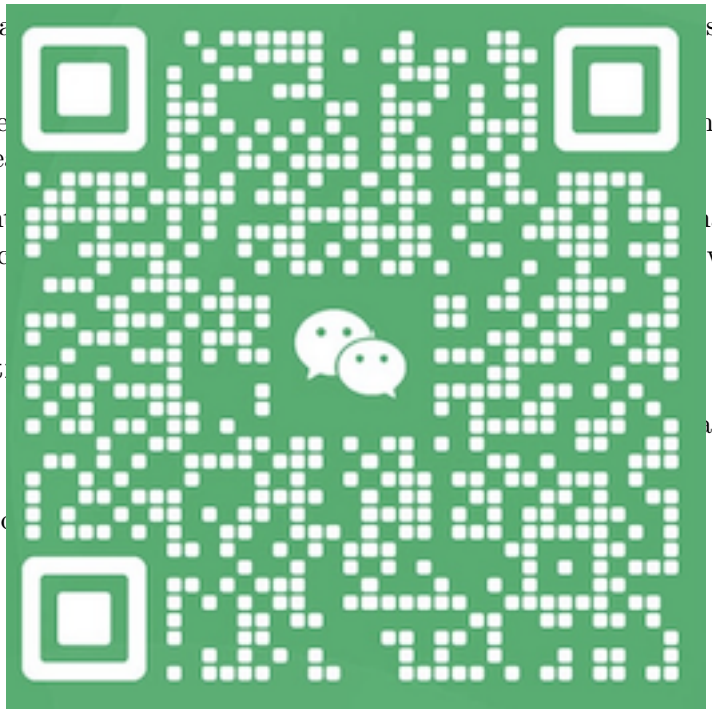
Question 1: General Concepts [24 marks]

a) Compare and contrast "lemmatisation" and "stemming". With the aid of an example word, show how lemmatisation and stemming differ. [6 marks]

b) What is "Relation Extra sumptions made about the set of relations. [6

c) Why is "topic model" e ethods, and also their shortcomings (one for e

d) "Copy mechanism" is in andle out-of-vocabulary words in the source doc words, justify whether copy mechanism is still



Question 2: Distributi

a) Compare and contrast arning "word vectors". [6 marks]

b) Explain one limitation

Continued overleaf . . .

Section B: Method Questions [46 marks]

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

Question 3: Hidden Markov Models [20 marks]

"Text chunking" refers to the preprocessing step where we divide a text into syntactically related non-overlapping groups of words. The following is an example of a chunked sentence:

```
[money] [could be added] [to] [a spending bill] [covering] [the federal emergency
management agency] , [which] [coordinates] [federal disaster relief] .
```

a) Show how a "Hidden Markov model" can be used to perform text chunking. In your answer, you should: (i) explain the state inventory; (ii) show how the given sentence would be labelled for training the HMM; (iii) describe how the model parameters can be learned (assuming we have a labelled corpus) and illustrate with examples from the given sentence. [12 marks]

b) Explain two drawbacks of using HMM for this task, and suggest a solution for each of the drawbacks [8 marks]

Question 4: Discourse

Consider the following par rithm described in the lecture to partition it into n_i" denotes the cosine similarity between the two ither side of gap i:

$\text{sim}_0 = 0.9$

Saturn i e Solar System.

$\text{sim}_1 = 0.3$

It is th after Jupiter.

$\text{sim}_2 = 0.1$

Saturn i word Saturnus, which me

$\text{sim}_3 = 0.2$

The Roma ity Kronos.

$\text{sim}_4 = 0.7$

Saturn h

The larg lanet Mercury.

a) Perform "discourse segmentation" on the paragraph, assuming $t = 0.2$. Your answer should show which gap a boundary will be inserted and the computation involved in producing the decision. [3 marks]

b) Discuss one drawback of using bag-of-words vectors for computing similarity, and illustrate how it created an erroneous boundary based on your answer in (a). You may assume that the bag-of-words vectors are created using "TF-IDF" weights with the following preprocessing steps: (1) symbols and numbers are removed; and (2) words are lowercased and tokenised using white space. [5 marks]

c) Propose 2 solutions that alleviate the drawback. With the aid of examples, show how the solutions will help improve discourse segmentation for the given paragraph. [6 marks]

Question 5: Ethics [12 marks]

Given the first application described in the guest lecture (automatic triaging of legal requests), discuss **three** ethical implications of this application.

Continued overleaf . . .

Section C: Algorithmic Questions [41 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 6: Context-free Grammar [23 marks]

Consider a "context-free grammar" with the following production rules:

```
S → NP VP NP          S → NP VP PP
NP → NP PP            NP → N
VP → VP NP            VP → V
PP → IN NP
N → he                N → elephants        N → trucks
V → loads
IN → onto
```

a) Convert the grammar int "Ch hat rules are modified, removed, and added. [

b) Given the sentence: `he` rsing" on it using the converted grammar. Y re, which includes the edges. Your solution sh

c) One of the parses is mor unconverted) grammar above has now been mo the parses. Propose a change to the probabilis er" to produce a higher probability for the mor nge, use parse trees to illustrate how the chang he parser producing a higher probability for t

Question 7: N-gram l

This question asks you to del". You should leave your answers as fractions. ence:

```
i wish to wish the
if you won't wish
i won't wish the wish you wish to wish
```

a) Compute the probability of all bigrams given the context word `wish` under a "unsmoothed bigram language model". [4 marks]

b) Compute the probability of all bigrams given the context word `wish` under a "bigram language model with absolute discounting" where the discount factor $d = 0.2$. [8 marks]

c) Compute the probability of **unseen bigrams** given the context word `wish` under a "bigram language model with Kneser-ney smoothing" where the discount factor $d = 0.2$. [6 marks]

— End of Exam —