Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

BSc Artificial Intelligence & Computer Science, BSc Artificial Intelligence & Computer Science with an Industrial Year, BSc Artificial Intelligence & Computer Science with Study Abroad , BSc Computer Science , BSc Computer Science with an Industrial Year, BSc Computer Science with Digital Technology Partnership (PwC), BSc Computer Science with Digital Technology Partnership (Vodafone), BSc Computer Science with Study Abroad, BSc Mathematics & Computer Science, BSc Mathematics & Computer Science with an Industrial Year, MEng Computer Science/Software Engineering, MEng Computer Science/Software Engineering with an Industrial Year, MEng Computer Science/Software Engineering, MEng Computer Science/Software Engineering with an Industrial Year, MRes Natural Computation, MSc Artificial Intelligence and Machine Learning, MSci Computer Science, MSci Computer Science with an Industrial Year, MSci Computer Science with Study Abroad, MSci Mathematics & Computer Science, MSci Mathematics & Computer Science with an Industrial Year

06-37810 / 06-37812

Natural Language Processing

Main Summer Examinations 2023

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 80, which will be rescaled to a mark out of 100.

Question 1

(a) How does inflectional and derivational morphology affect the computational performance of NLP applications? Give an example of this issue in relation to a particular application in your answer. Be clear to distinguish inflectional from derivational morphology. **[5 marks]**

(b) Propose two techniques to attempt to deal with the issues with morphology you have identified in part (a). Describe how the techniques could be implemented and how their performance could be tested. **[5 marks]**

(c) "The more often term x is used in a document, the more relevant that document becomes to a query containing x amongst its query terms."

Discuss the degree to which you agree with this statement. **[5 marks]**

(d) What are the advantages and disadvantages of using Mean Reciprocal Rank as an evaluation method for an Information Retrieval system? What are the risks of solely using this as an evaluation method? **[5 marks]**

Question 2

(a) Give a definition of Mutual Information in the context of the text classification pipeline and explain its role. **[5 marks]**

(b) What makes a Naïve Bayes classifier naïve? Given a practical example for text classification of overcoming this naïvety. **[5 marks]**

(c) Maximum Likelihood training in Naïve Bayes for text classification is problematic. Describe the problem and give a practical solution. **[5 marks]**

(d) Assume the following likelihoods from Table 1 (on the following page) for each word being part of a positive or negative book review and equal prior probabilities for each class.

Show, with your workings, what class Naïve Bayes will assign to the sentence "Its redeeming strength is authenticity."

[5 marks]

	Positive	Negative
Its	0.01	0.2
redeeming	0.1	0.001
strength	0.1	0.01
weakness	0.003	0.3
is	0.05	0.005
authenticity	0.1	0.01

Table 1: Likelihood values

Question 3

Consider the following "document", where Word1 to Word6 are the words in the vocabulary and each line (enclosed by brackets) represents a sentence:

[Word3 Word3 Word6 Word6 Word2] (1)
[Word1 Word2 Word3 Word5 Word2 Word4] (2)
[Word4 Word6 Word6 Word4 Word6] (3)
[Word2 Word6 Word5 Word1] (4)
[Word3 Word6 Word3 Word2 Word1] (5)

(a) Given the above "document", obtain a (distributional) semantic representation of each word, based on co-occurrence. Write down the full co-occurrence matrix for the document, following the rules: **[8 marks]**

- Two words co-occur if they appear in the same sentence
- Ignore multiple co-occurrences of the same word pair within the same sentence
- Co-occurrences of the same word pairs in different sentences count separately

(b) Calculate the cosine similarity between Word1 and each other word in the corpus. Which word is most similar to Word1? **[6 marks]**

(c) Are there other ways to measure similarity in distributional space? Can you explain how they would work? With reference to the methods you have presented, outline the key differences between the method and cosine similarity? **[6 marks]**

Question 4

Intelligently discuss the components required to build a system for Automatic Fact-Checking. This discussion should be no more than 2 pages in length. The discussion should be structured around the NLP pipeline and how this framework could be applied to the task of Automatic Fact-Checking. **[20 marks]**

End of Paper

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.