A37937

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

Third Year Undergraduate

06-37810

LH Natural Language Processing

Main Summer Examinations 2025

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions. Questions 1 and 2 are worth 30 marks each, and Question 3 is worth 40 marks. The paper will be marked out of 100.

Question 1

(a) Calculate the minimum edit distance, and show the calculations and results in a relevant table, between the words "apfel" and "apple" using the Levenshtein distance cost in the process. **[10 marks]**

(b) Given the following corpus:

- "the dog barked"
- "the dog ran away"
- "the cat meowed"
- "the cat and the dog played"

Calculate the bigram probabilities for the sentence "the dog played". Use Laplace smoothing and show your workings.

[10 marks]

(c) Interpolation and backoff are approaches used in language modeling. Define each and argue for why one approach is preferable to the other. **[5 marks]**

(d) Why and how would you intrinsically evaluate a language model? **[5 marks]**

Question 2

(a) Word2vec and ELMO are two popular frameworks for obtaining word embeddings. Consider the following two sentences:

- She has a blue car.
- He was feeling blue.

What is the key difference in the representations that we would obtain from Word2Vec and ELMO for the word "blue" in these sentences? What type of neural architecture do Word2Vec and ELMO use to obtain the representations? **[5 marks]**

(b) Consider the following six words and their vector representations:

blue: [6,7,2,9]
house: [8,3,4,2]
cat: [1,2,3,4]
car: [4,8,1,3]
red: [9,5,2,7]
and: [3,5,6,1]

Describe at least two different ways to combine word vectors into vectors for phrases and sentences. Discuss the advantages and disadvantages of each of them. Pick one strategy and calculate a vector representation for the following phrases: "blue house", "blue cat", "red car", "red house and blue cat".

[5 marks]

(c) Transformer neural networks are typically grouped in three different types: encoders, decoders, and encoder-decoders. Can you give an example for an encoder transformer, a decoder transformer, and an encoder-decoder? What is the training objective for the encoder and for the decoder and how are they different? If you want to use a transformer model for generating free text, which of the encoder and decoder will perform better? **[10 marks]**

(d) During "reinforcement learning from human feedback", large language models are interacting with a "reward model". What is the purpose of the reward model?; what type of machine learning model is the reward model (supervised, unsupervised, reinforcement)?; how is it trained? **[10 marks]**

Question 3

Consider a hypothetical scenario where you are tasked with developing an NLP model to understand and generate song lyrics. Propose an approach to tackle this problem, discussing practical issues such as the data involved, preprocessing techniques, model choice, model implementation and how to evaluate your system. You should also consider and discuss aspects such as text structure, rhymes, and meter (repeating patterns in the lyrics). What challenges do you anticipate, and how might you address them? Your answer should span no more than two pages. **[40 marks]**

This page intentionally left blank.

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.