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

Natural Language Processing

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Turn Over

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60.

Question 1

(a) Write a regular expression to match a line of text that starts with a word where the first letter is capitalised and the line ends with either a smiley emoticon **:-)** or a sad face emoticon **:-(**

[5 marks]

(b) Given the following whitespace-separated list of character sequences, apply the Byte Pair Encoding Token Learner algorithm for 3 passes over this, showing results at each pass:

:-) :-) :-) :-) :-) happy happy happy happy hapy :-(:-(:-(risky risky risky risky

[5 marks]

(c) Given the following short, restaurant reviews, each labelled with a sentiment of either **positive** or **negative** :

Document	Class
Really great, really great :-)	positive
Really great ambiance and experience :-)	positive
Terrible restaurant	negative
Not really as good as I expected :-(	negative
Food was cold :-(	negative

Calculate how a Binary Multinomial Naïve Bayes classifier with Laplace smoothing would classify the following document: "Really cracking food B-)"

[5 marks]

(d) Having worked with the restaurant review data you have a feeling that there are some fake review documents in amongst the real reviews. Discuss the effects that this would have when evaluating your text classification system with respect to standard metrics. **[5 marks]**

Question 2

(a) Word2vec is a popular framework for obtaining word embeddings. Answer the following questions about word2vec: 1) what are word embeddings; 2) what is the difference between the two word2vec algorithms: CBOW and SKIP-GRAM; 3) how does SKIP-GRAM measure similarity between words? **[5 marks]**

(b) You are given the following 6 words and their 4 dimensional vector representations. Calculate the similarity between "computer" and each of the other 5 words. Order the words from the most similar to the least similar.

computer: [6,7,2,9]
house: [8,3,4,2]
cat: [1,2,3,4]
car: [4,8,1,3]
mouse: [9,5,2,7]
swimming: [3,5,6,1]

[5 marks]

(c) You are given the sentence "the quick brown fox jumped over the lazy dog". Consider three different models trained to assign probability of words: 1) a uni-directional forward LSTM; 2) an encoder transformer (BERT), trained with masking a single word; 3) a decoder transformer (GPT). For each of the three models, identify the context words that will be used to calculate the probability of seeing the word "fox". Justify your answer by discussing how each model calculated the probability of words.
[5 marks]

(d) We discussed two approaches to solving NLP problems: using a machine learning pipeline and using end-to-end neural network. Explain what is an end-to-end approach and list at least two advantages and disadvantages compared to a traditional pipeline approach. **[5 marks]**

Question 3

Consider a hypothetical scenario where you are tasked with developing an NLP model to understand and generate poetic text. 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 like text structure, rhymes, and meter and how your approaches tackles these. What challenges do you anticipate, and how might you address them? Your answer should span no more than two pages. **[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.