EXAMINATIONS – 2025

TRIMESTER TWO

AIML 332

AI Natural Language Processing

Time Allowed: 120 MINUTES

CLOSED BOOK

Permitted materials: Only silent non-programmable calculators or silent programmable calculators with their memories cleared are permitted in this test.
Non-electronic foreign language translation dictionaries can be used.

Instructions: There are a total of 100 marks on this test.
Attempt all questions.

Questions

Questions

1. Text representation [13]
2. Text classification [17]
3. Attention and transformers [25]
4. LLM alignment [10]
5. Prompt expansion mechanisms [10]
6. Recent and future developments in LLMs [15]
7. Web search and recommender systems [10]

1. Text representation **(13 marks)**

(a) TF-IDF **(9 marks)**

Consider a corpus containing three sentences:

Sentence 1: "The cat sat on the mat."

Sentence 2: "The dog sat on the log."

Sentence 3: "The cat chased the dog."

(i) Show all the features/terms in the corpus (ignore capitalization and punctuation).

(ii) Compute the IDF values for **each feature/term**. The inverse document frequency (IDF) of a term t is defined as:

$$\text{IDF}(t) = 1 + \log\left(\frac{N+1}{\text{df}(t)+1}\right)$$

(iii) Compute the TF-IDF vector for **Sentence 1 only**. The normalization step is not required.

(b) Briefly state two key differences between CBOW and Skip-gram in Word2Vec? **(4 marks)**

2. Text classification **(17 marks)**

(a) Briefly explain how pre-trained word embeddings are typically used in NLP models. State one advantage and one disadvantage of using pre-trained word embeddings. **(6 marks)**

(b) A CNN uses filters of size 3, 4, and 5 over an input sentence represented by word embeddings. What kinds of textual features are these filters detecting? **(3 marks)**

(c) In a CNN-based model for text classification, different layers contain different types of learnable parameters. Briefly describe the parameters in (i) the embedding layer, (ii) convolutional layers, and (iii) fully connected layers (iv) any other optional layers **(8 marks)**

3. Attention and transformers **(25 marks)**

The **GPT-2** transformer is set up as a **decoder** network, that learns to generate one token (at a time), from an input sequence of tokens. The decoder network implements a **self-attention** mechanism, that learns to map tokens in the input sequence 'back to themselves'.

(a) The attention mechanism represents each input token t as a *weighted sum* of other input tokens (in some embedding space). Explain why, for a decoder network like GPT-2, this weighted sum does not include tokens *to the right* of t. **(4 marks)**

(b) GPT-2's self-attention mechanism learns *patterns* in the input sequence: that is, commonly-occurring *combinations* of input tokens, which may be far apart in the sequence. Explain why self-attention is better at learning linguistic patterns than a recurrent neural network (RNN). Your answer should include a brief description of how a RNN works, and should mention the 'bottleneck problem'. **(6 marks)**

(c) To compute the weights for its weighted sum, the attention mechanism encodes the input sequence as two separate matrices of **token embeddings**, Q and K. Q represents tokens as 'items to be summed'; K represents each token as a 'target to be matched'. The core matrix multiplication is given by $\underline{QK^T}$ (where K^T is the transpose of K).

i. QK^T computes the **distance** between each target token embedding t and each input item embedding i. Explain how distances are represented, and how matrix multiplication computes them. **(3 marks)**

ii. To convert the matrix of distances computed by QK^T into a matrix of weights, we apply a **masking** operation, then a **softmax** operation. Explain how these two operations work, for GPT-2. **(4 marks)**

iii. The operations that compute the weights matrix (QK^T, masking and softmax) run mechanically: there are no learnable parameters. The mechanism responsible for learning 'the right weights' operates earlier in the attention system, in the networks that map tokens to points in embedding space. Explain what these networks do, and how their parameters are adjusted during training. **(4 marks)**

iv. Questions (i)–(iii) describe a *single* attention mechanism. But GPT-2 actually implements *several independent* attention mechanisms, referred to as separate **attentional heads**. Explain why it's useful to have several heads, for learning linguistic patterns in the input sequence. **(4 marks)**

4. LLM alignment **(10 marks)**

After a LLM is pretrained, on a huge corpus of text, it is **aligned** on smaller, more carefully curated training datasets.

(a) One alignment process is needed to turn a general **language model** into a **dialogue model**. What form does the training data take for this process? Illustrate with an example. **(4 marks)**

(b) Other alignment processes are needed to encourage the dialogue model to produce responses that are 'helpful', and not 'harmful'. Describe two different technical mechanisms that can be used to implement this kind of alignment, with an example of each. **(6 marks)**

5. Prompt expansion mechanisms **(10 marks)**

(a) Explain how **retrieval-augmented generation** (RAG) expands a user LLM prompt, to produce a better response. **(5 marks)**

(b) RAG exploits the principle that a LLM generally performs better when its prompt is *longer* and *more informative*. Why is this the case? Your should make reference to LLMs' ability to learn probabilities conditioned on long input sequences, and to generalise away from their training inputs. **(5 marks)**

6. Recent and future developments in LLMs **(15 marks)**

(a) Using LLM APIs, programmers can write simple code routines that call LLMs iteratively, or conditionally (sometimes called 'LLM workflows'). Give an example of one of these routines, and explain its use. You may describe the routine with pseudocode, or in a flowchart diagram. **(5 marks)**

(b) Representations of text learned by LLMs can be used in 'multimodal' systems that operate with text and images - for instance, mapping images to text, or text to images. Describe one such system, and explain how it makes use of a LLM. **(5 marks)**

(c) Describe one way the performance of an LLM can be improved by 'scaling' some aspect of its operation. Can scaling continue indefinitely, for your chosen method? Briefly explain your answer.
(5 marks)

7. Web search and recommender systems **(10 marks)**

(a) Compare and contrast collaborative filtering and content-based filtering in terms of data requirements, strengths, and weaknesses. **(6 marks)**

(b) Describe two advantages and two challenges of using large language models (LLMs) for search and information retrieval. **(4 marks)**

* * * * * * * * * * * * * * *

Student ID: .....

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.

Student ID: .....

SPARE PAGE FOR EXTRA ANSWERS

Cross out rough working that you do not want marked.
Specify the question number for work that you do want marked.