VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

EXAMINATIONS – 2025

TRIMESTER 1

AIML428

Text Mining and NLP

Time Allowed: TWO HOURS

CLOSED BOOK

Permitted materials: Only silent non-programmable calculators or silent programmable calculators with their memories cleared are permitted in this examination.

Non-electronic foreign/English language dictionaries are permitted.

Instructions: Answer all questions.

The exam will be marked out of 100.

Please do not write long paragraphs. Use lists if possible.

Questions

1.	Text representation	[20]
2.	Text classification	[10]
3.	Attention and transformers	[15]
4.	Language models and tasks	[15]
5.	Language model alignment	[15]
6.	Chain-of-thought AI systems and LLM APIs	[15]
7.	Web search	[5]
8.	Recommender systems	[5]

1. Text representation **(20 marks)**

(a) TF-IDF **(10 marks)**

Given a small corpus with three short documents:

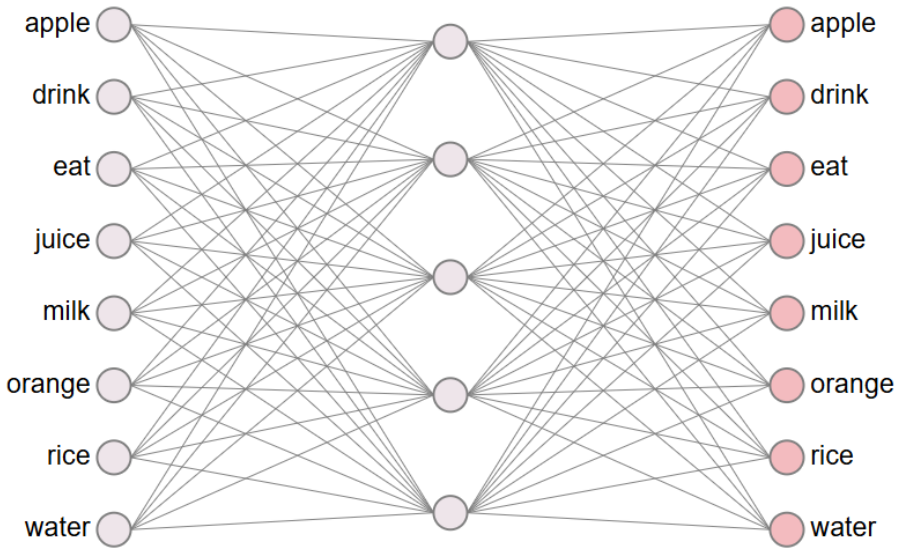
- Fight fire with fire
- Can you CAN a can as a canner can CAN a can?
- You can do it

Compute and present the TF-IDF representation for each document. Please note that the final normalization step is not needed.

Choose two representative TF-IDF values and illustrate the calculation process, including all intermediate steps: TF, DF, IDF, and the resulting TF-IDF weights.

(b) Word embedding **(10 marks)**

Suppose you are using the following small neural network to explain how word embedding can be learned using the CBOW model in Word2Vec. Please note this is a toy example with only eight words in the vocabulary, and a context window size of just one word.



i. Give two training examples, and show the one-hot encoding of the input and output of each training example.

ii. Use a single word from one of your training examples to explain where to obtain the embedding values from the trained network. You may do this by drawing the portion of the network relating to your example word, highlighting and labelling the connections/weights that become its embedding representation.

2. Text classification **(10 marks)**

(a) Evaluation criteria **(5 marks)**

The following is the confusion matrix of a binary text classifications system.

		ACTUAL: Negative	ACTUAL: Positive
PREDICTION	Negative	60	8
	Positive	22	10

Calculate the overall accuracy of the system and the precision, recall, F1 measure of the Positive class. Please show your working.

(b) Comparing a CNN for text classification and a CNN for image classification, what is the main difference in the convolutional layer? Briefly explain why. **(5 marks)**

3. Attention and transformers **(15 marks)**

In modern language models, **attention** is used to learn a function from input words to 'the next word'.

(a) Describe two key properties of attention that make it a more efficient learning mechanism than a recurrent network. **(5 marks)**

(b) In a **transformer**, attention also learns to represent structure in the sequence of input words by themselves. Explain how it does this. **(5 marks)**

(c) A transformer maps a sequence of input words to a whole *sequence* of output words. Another attention mechanism learns the structure of this output sequence. Explain how this attention mechanism differs from the one applied to the input. **(5 marks)**

4. Language models and tasks **(15 marks)**

A transformer learns to perform a **sequence-to-sequence task**, mapping an input word sequence to an output sequence.

(a) Describe two sequence-to-sequence tasks, in different domains of language processing. Illustrate each with an example. **(3 marks)**

(b) Early language models were trained for specific tasks, but modern transformers learn *task-general* language models. Explain how modern models acquire these task-general abilities. **(4 marks)**

(c) Modern transformers also let the user give details of a specific task to a trained system *in its prompt*, using 'few-shot learning'.

i. Give an example of a prompt that uses few-shot learning to guide the transformer. **(4 marks)**

ii. Explain how few-shot learning encourages the transformer to deliver a task-specific response. What kind of learning is in play here? **(4 marks)**

5. Language model alignment **(15 marks)**

After they are trained, modern language models are **aligned**, to encourage safe responses.

(a) Explain how the alignment process differs from the initial training process. **(5 marks)**

(b) What kind of training data is used in the alignment process? Illustrate with an example training item. **(5 marks)**

(c) What kind of training is used for the alignment process you illustrated? **(5 marks)**

6. Chain-of-thought AI systems and LLM APIs **(15 marks)**

While a LLM responds directly to a user instruction, a **chain-of-thought AI system** is a program that takes a user instruction and executes a *sequence* of LLM calls, to deliver a response constructed in several steps.

(a) Give an example of a user instruction which would plausibly be better handled by a chain-of-thought AI system than by a single LLM call. Explain why you think a chain-of-thought system would provide a better response. **(5 marks)**

(b) Write some pseudocode implementing a simple chain-of-thought for your chosen user instruction. Your chain of thought can include agentic actions, if you wish. **(10 marks)**

7. Web search **(5 marks)**

In your opinion, what should be the main characteristics of the next generation of web search engines?

8. Recommender systems **(5 marks)**

Suppose you're developing a recommender system for a new database containing large amounts of text files, some audio and video files. Specify a good recommender algorithm to start with and a more advanced recommender algorithm for long-term use. Justify your answer.

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