EBU4203 Introduction to AI – Week 4 Tutorial 2023

Q1: This question is about Natural Language Processing (NLP).

a) What is Natural Language Processing (NLP), and why is it an essential field in artificial intelligence and linguistics?
b) Provide examples of real-world NLP applications.
c) Tokenization and Stemming are two commonly used text preprocessing techniques in NLP. Explain the functionality of these two techniques.

Q2. You are working with a small text corpus containing four sentences:

i. "Word embeddings are essential."
ii. "Word2Vec is a popular technique."
iii. "NLP tasks benefit from word embeddings."
iv. "Word2Vec models capture word similarities."

Perform the following tasks related to Word2Vec:

1) **Tokenization:** Tokenize each sentence into individual words and list them.
2) **Vocabulary Size:** Calculate the total number of unique words in the corpus.
3) **Word Analogy:** If the vector for "king" - "man" + "woman" results in a new vector, what concept or word might be represented by this new vector? Provide a brief explanation.

Q3. Consider two word vectors:

- Vector A: [0.6, 0.8]
- Vector B: [0.3, 0.4]

Calculate the cosine similarity between vectors A and B. Show all calculations and provide the cosine similarity score.

Q4. What are the limitations of current AI approach?

Q5. What is domain shift? Give three solutions to release domain shift.

Q6. Give three examples of Artificial generative intelligence (AGI). How will AGI evolve in the future?

Q7. Consider a simplified Hidden Markov Model (HMM)-based part-of-speech tagging system with three part-of-speech tags: Noun (N), Verb (V), and Model (M). There are three sentences in the training set

1. Time will fly.

2. Will he cook?

3. Will can cook.

- **Table 1** shows the table of probabilities of each word appeared as respective part-of-speech tag. Fill in the missing values labelled by "?" in **Table 1.**

Words	**Noun**	**Model**	**Verb**
time	?	0	0
will	?	?	?
fly	?	0	?
he	?	0	0
cook	0	0	?
can	0	?	0

Table 1

- Table 2 shows the co-occurrence table, which can be used to analyze how different parts-of-speech interact within a text corpus. Fill in the missing values labelled by "?" in Table 2.

	Noun	Model	Verb	<End>
<Start>	2	?	?	?
Noun	?	?	?	?
Model	?	?	?	0
Verb	?	?	?	?

Table 2

Q8 Consider the following sentence

"The curious cat quietly approached the mysterious door."

You are required to develop the training set for generating word embedding. When the index of the centre word is $i = 2$ and the window size is $W = 2$, the source text and training samples will look like this

Source Text

Training Samples

➡ (?, ?)

The	curious	**cat**	quietly	approached	the	mysterious	door	

1) Define the training set when $i = 2, W = 1$
2) Define the training set when $i = 1, W = 2$
3) Define the training set when $i = 0, W = 6$