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?
Tips: Slides: p38-39
Natural Language Processing (NLP) is a vital field in artificial intelligence and linguistics that empowers computers to understand, interpret, and generate human language, facilitating more natural communication between humans and machines. Its applications range from automating tasks and data understanding to multilingual communication and personalized content creation, making NLP an essential component of modern technology and information processing.

b) Provide examples of real-world NLP applications.
Tips: Slides: p41
NLP has practical applications in sentiment analysis, machine translation, speech recognition, and chatbots, among others.

Tokenization and Stemming are two commonly used text preprocessing techniques in NLP. Explain the functionality of these two techniques.
Tips: Slides: p50 -52
Tokenization: Splitting text into words or subword units.
Stemming: Reducing words to their base or root form.

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.

Tips: Slides: p50

Explanation： Tokenization is the process of breaking text into individual words or tokens. In this subtask, we divided each of the four sentences into their constituent words. This step is essential for further text analysis in NLP.

i. Sentence 1: ["Word", "embeddings", "are", "essential."]
ii. Sentence 2: ["Word2Vec", "is", "a", "popular", "technique."]
iii. Sentence 3: ["NLP", "tasks", "benefit", "from", "word", "embeddings."]
iv. Sentence 4: ["Word2Vec", "models", "capture", "word", "similarities."]

2) **Vocabulary Size:** Calculate the total number of unique words in the corpus.

Explanation： The vocabulary size represents the total number of **unique** words in the dataset.

The total number of unique words in the corpus is 19.

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.

Tips: Slides: p63

The concept or word represented by the vector resulting from "king" - "man" + "woman" is often associated with "queen."

This is a well-known analogy in Word2Vec embeddings, demonstrating that vector arithmetic can capture relationships between words. It represents the idea of changing the gender while maintaining a similar semantic relationship.

So, "queen" might be represented by the new vector.

Explanation: The vector "king" - "man" represents the gender relationship, and adding "woman" to it should result in a word that represents the female counterpart of "king," which is "queen."

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.

Tips: Slides: p65

Answer:

- Cosine Similarity ($\cos\theta$) = (A · B) / (||A|| * ||B||)
- Calculate dot product (A · B) = (0.6 * 0.3) + (0.8 * 0.4) = 0.18 + 0.32 = 0.50
- Calculate magnitude ||A|| = √(0.6^2 + 0.8^2) = √(0.36 + 0.64) = √1.00 = 1.00
- Calculate magnitude ||B|| = √(0.3^2 + 0.4^2) = √(0.09 + 0.16) = √0.25 = 0.50
- Cosine Similarity ($\cos\theta$) = (0.50) / (1.00 * 0.50) = 1.0

Q4. What are the limitations of current AI approach?

Tips: Slides: p122-127

Data dependency, Interpretability and Explainability, generalization, Computation and Resource Requirements, Energy Consumption, Robustness and Security.

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

Tips: Slides: p124

Domain shift refers to a phenomenon in machine learning and statistics where the statistical properties of data change when transitioning from one domain or distribution to another. In other words, it occurs when a model that is trained on data from one source domain performs poorly when applied to a different target domain.

Solutions: zero-shot learning, knowledge transfer, GAN.

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

Tips: Slides: p129

Examples: AlphaGo, ChatGPT, AlphaFold, ClimaX.....

Possible evolution direction: from common to professional, more media will be involved, LLMs leverage tools to affect real world, automatic driving, natural language will be the new programming language, AI for science.....

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.

1) **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

Tips: Slides: p33

Explanation: i.e. the word Will appears one time as a noun and 2 times as a model. To calculate the emission probabilities, create a counting table in a similar manner. The probabilities of Will as a noun takes 1/3, where 3 = 1 (time)+ 1(Will)+1(he).

Answer:

Words	Noun	Model	Verb
time	1/3	0	0
will	1/3	2/3	0
fly	0	0	1/3
he	1/3	0	0
cook	0	0	2/3
can	0	1/3	0

2) 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.

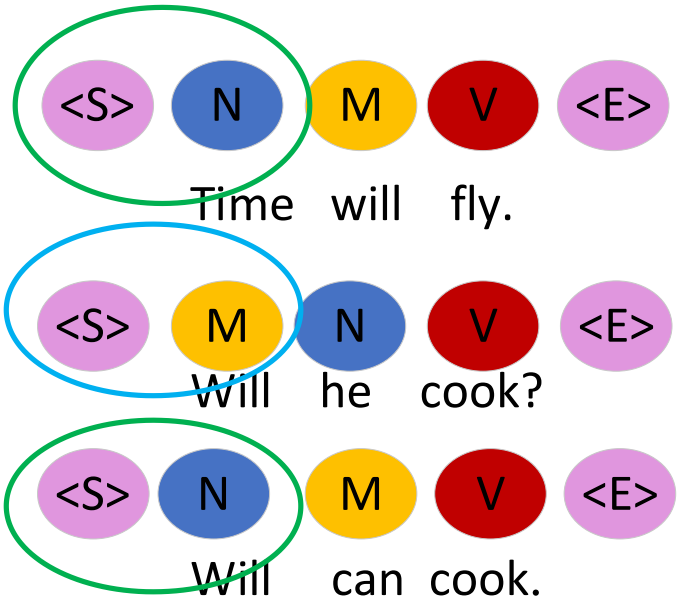
	Noun	Model	Verb	<End>

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

Table 2

Tips: Slides: p36

Explanation : fill it with the co-occurrence counts of the tags.



Answer;

	Noun	**Model**	**Verb**	**<End>**
<Start>	2	1	0	0
Noun	0	2	1	0
Model	1	0	2	0
Verb	0	0	0	3

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 $i = 0, W = 6$

Tips: Slides: p51

Answer;

i)

(cat, curious)

(cat, quietly)

ii)

(curious, The)

(curious, cat)

(curious, quietly)

iii)

(The, curious)

(The, cat)

(The, quietly)

(The, approached)

(The, the)

(The, mysterious)

(The, door)