Introduction to Large Language Models
Assignment- 4

Number of questions: 7   **Total mark: 6 X 1 + 1 X 4 = 10**

_____

QUESTION 1: [1 mark]

What is the main drawback of representing words as one-hot vectors?
 a. They cannot capture semantic similarity between words.
 b. They are computationally inefficient.
 c. They cannot incorporate word order effectively.
 d. They are not robust to unseen words.

Correct Answer: a
Solution: One-hot vectors are orthogonal and do not capture relationships or similarities between words.

_____

QUESTION 2: [1 mark]

What is the key concept underlying Word2Vec?
a. Ontological semantics
b. Decompositional semantics
c. Distributional semantics
d. Morphological analysis

Correct Answer: c
Solution: Please refer to lecture slides.

_____

QUESTION 3: [1 mark]

Why is sub-sampling frequent words beneficial in Word2Vec?
 a. It increases the computational cost.
 b. It helps reduce the noise from high-frequency words.
 c. It helps eliminate redundancy.
 d. It prevents the model from learning embeddings for common words.

Correct Answer: b
Solution: Sub-sampling helps reduce the effect of very frequent but less informative words.

_____

QUESTION 4:
Which word relations cannot be captured by word2vec? **[1 mark]**
a. Polysemy
b. Antonymy

c. Analogy
d. All of the these

Correct Answer: a,b
Solution: word2vec vectors cannot effectively handle word relations like antonymy or polysemy.

_____

For Question 5 to 6, Consider the following word-word matrix:

	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}
w_1	1	5	3	0	1	5	7
w_2	4	2	4	1	6	2	0
w_3	2	1	9	2	5	1	5
w_4	5	0	7	4	2	0	4
w_5	3	5	1	0	1	2	1

QUESTION 5: [1 mark]
Compute the cosine similarity between w_2 and w_5.

a. 0.516
b. 0.881
c. 0.705
d. 0.641

Correct Answer: d
Solution:

$$\text{cos-sim}(\mathbf{w_2}, \mathbf{w_5}) = \frac{w_2 \cdot w_5}{||w_2|| \, ||w_5||}$$

$$\text{cos-sim}(\mathbf{w_2}, \mathbf{w_5}) = \frac{4*3 + 2*5 + 4*1 + 1*0 + 6*1 + 2*2 + 0*1}{\sqrt{4^2 + 2^2 + 4^2 + 1^2 + 6^2 + 2^2 + 0^2} \, \sqrt{3^2 + 5^2 + 1^2 + 0^2 + 1^2 + 2^2 + 1^2}}$$

$$= \frac{36}{\sqrt{77} \, \sqrt{41}} = 0.6407$$

QUESTION 6: [4 mark]
Which word is most similar to w_1 based on cosine similarity?
a. w_2
b. w_3
c. w_4
d. w_5

Correct Answer: d
Solution:

$$\text{cos-sim } (\mathbf{w_1}, \mathbf{w_2}) = \frac{1*4 + 5*2 + 3*4 + 0*1 + 1*6 + 5*2 + 7*0}{\sqrt{1^2 + 5^2 + 3^2 + 0^2 + 1^2 + 5^2 + 7^2} \ \sqrt{4^2 + 2^2 + 4^2 + 1^2 + 6^2 + 2^2 + 0^2}}$$

$$= \frac{42}{\sqrt{110} \ \sqrt{77}} = 0.45636$$

$$\text{cos-sim } (\mathbf{w_1}, \mathbf{w_2}) = \frac{1*2 + 5*1 + 3*9 + 0*2 + 1*5 + 5*1 + 7*5}{\sqrt{1^2 + 5^2 + 3^2 + 0^2 + 1^2 + 5^2 + 7^2} \ \sqrt{2^2 + 1^2 + 9^2 + 2^2 + 5^2 + 1^2 + 5^2}}$$

$$= \frac{79}{\sqrt{110} \ \sqrt{141}} = 0.63434$$

$$\text{cos-sim } (\mathbf{w_1}, \mathbf{w_2}) = \frac{1*5 + 5*0 + 3*7 + 0*4 + 1*2 + 5*0 + 7*4}{\sqrt{1^2 + 5^2 + 3^2 + 0^2 + 1^2 + 5^2 + 7^2} \ \sqrt{5^2 + 0^2 + 7^2 + 4^2 + 2^2 + 0^2 + 4^2}}$$

$$= \frac{56}{\sqrt{110} \ \sqrt{110}} = 0.5091$$

$$\text{cos-sim } (\mathbf{w_1}, \mathbf{w_2}) = \frac{1*3 + 5*5 + 3*1 + 0*0 + 1*1 + 5*2 + 7*1}{\sqrt{1^2 + 5^2 + 3^2 + 0^2 + 1^2 + 5^2 + 7^2} \ \sqrt{3^2 + 5^2 + 1^2 + 0^2 + 1^2 + 2^2 + 1^2}}$$

$$= \frac{49}{\sqrt{110} \ \sqrt{41}} = 0.7296$$

QUESTION 7: [1 mark]

What is the difference between CBOW and Skip-Gram in Word2Vec?
 a. CBOW predicts the context word given the target word, while Skip-Gram predicts the target word given the context words.
 b. CBOW predicts the target word given the context words, while Skip-Gram predicts the context words given the target word.
 c. CBOW is used for generating word vectors, while Skip-Gram is not.
 d. Skip-Gram uses a thesaurus, while CBOW does not.

Correct Answer: b
Solution: CBOW predicts the middle word in a context, whereas Skip-Gram predicts the surrounding words for a given target word.