Natural Language Processing

Assignment 7

Type of Question: MCQ

Number of Questions: 7 **Total Marks: (5×1)+(3×2)=10**

==

Question 1: Suppose you have a raw text corpus and you compute word co occurrence matrix from there. Which of the following algorithm(s) can you utilize to learn word representations? (Choose all that apply) **[1 mark]**

a. CBOW
b. SVD
c. PCA
d. GloVe

Answer: a, b, c, d
Solution:

==

Question 2: What is the method for solving word analogy questions like, given A, B and D, find C such that A:B::C:D, using word vectors? **[1 mark]**

a. $v_c = v_a + (v_b - v_d)$, then use cosine similarity to find the closest word of v_c. b.
$v_c = v_a + (v_d - v_b)$ then do dictionary lookup for v_c
c. $v_c = v_d + (v_b - v_a)$ then use cosine similarity to find the closest word of v_c. d.
$v_c = v_d + (v_a - v_b)$ then do dictionary lookup for v_c.
e. None of the above

Answer: e

Solution: $v_d - v_c = v_b - v_a$

$v_c = v_d + v_a - v_b$ then use cosine similarity to find the closest word of v_c.

==

Question 3: What is the value of $PMI(w_1, w_2)$ for $C(w_1) = 100$, $C(w_2) = 2000$, $C(w_1,$

w_2) = 64, N = 100000? N: Total number of documents.
$C(w_i)$: Number of documents, w_i has appeared in.
$C(w_i, w_j)$: Number of documents where both the words have appeared in. Note: Use base 2 in logarithm. **[1 mark]**

a. 4
b. 5
c. 6
d. 5.64

Answer: b

Solution:

$$PMI = \log_2 \frac{64 \times 100000}{100 \times 2000} = 5$$

==

Question 4: Given two binary word vectors w_1 and w_2 as follows:
w_1 = [1010101010]
w_2 = [0011111100]
Compute the Dice and Jaccard similarity between them. **[2 marks]**

a. 6/11, 3/8
b. 10/11, 5/6
c. 4/9, 2/7
d. 5/9, 5/8

Answer: a

$$\text{Dice coefficient} = \frac{2 \times 3}{5 + 6} = \frac{6}{11}$$

$$\text{Jaccard coefficient} = \frac{3}{8}$$

Solution:

==

Question 5: Consider two probability distributions for two words be p and q. Compute their similarity scores with KL-divergence. **[2mark]** p = [0.20, 0.75, 0.50]
q = [0.90, 0.10, 0.25]
Note: Use base 2 in logarithm.

a. 4.704, 1,720
b. 1.692, 0.553
c. 2.246, 1.412
d. 3.213, 2.426

Answer: c
Solution:

$$\begin{aligned}\text{KL-div}(p,q) &= \sum_i p_i \log_2 \frac{p_i}{q_i} \\ &= 0.2\log\frac{0.2}{0.9} + 0.75\log\frac{0.75}{0.1} + 0.5\log\frac{0.5}{0.25} \\ &\approx 2.246 \\ \text{KL-div}(q,p) &= 0.9\log\frac{0.9}{0.2} + 0.1\log\frac{0.1}{0.75} + 0.25\log\frac{0.25}{0.5} \\ &\approx 1.412\end{aligned}$$

==

Question 6: Consider the following word co-occurrence matrix given below. Compute the cosine similarity between
(i) w1 and w2, and (ii) w1 and w3. **[2 mark]**

	w4	w5	w6
w1	2	9	4
W2	1	5	6
W3	3	0	1

a. 0.773, 0.412
b. 0.881, 0.764
c. 0.665, 0.601
d. 0.897, 0.315

Answer: d
Solution:

$$\text{cosine-sim}\ (\vec{p}, \vec{q}) = \frac{\vec{p} \cdot \vec{q}}{\|\vec{p}\| \cdot \|\vec{q}\|}$$

$$\text{cosine-sim}\ (\text{w1, w2}) = \frac{2 \times 1 + 9 \times 5 + 4 \times 6}{\sqrt{2^2 + 9^2 + 4^2} \times \sqrt{1^2 + 5^2 + 6^2}} \approx 0.897$$

$$\text{cosine-sim}\ (\text{w1, w3}) \approx 0.315$$

==

Question 7: Which of the following statement(s) is/are True? **[1 mark]**

a. In structured distributional semantics, co-occurrence statistics are collected using parser extracted relations.
b. Term mismatch occurs from the word independence assumption during document indexing.
c. We can use distribution semantic models for query expansion.
d. Attributional similarity depends on the degree of correspondence between attributes.

Answer: a, b, c, d
Solution: