Natural Language Processing

Assignment 7

Type of Question: MCQ

Number of Questions: 7

==========

Question 1: **[1 mark]**

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)

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

Answer: a, b, c, d
Solution:

==========

Question 2: **[1 mark]**

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?

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. $v_c = v_d + (v_a - v_b)$ 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: c
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: **[1 mark]**

What is the value of $PMI(w_1, w_2)$ for $C(w_1) = 100$, $C(w_2) = 2500$, $C(w_1, w_2) = 320$, $N = 50000$? 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.

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

Answer: c

Solution:

PMI = log2 [(320*50000) / (100*2500)] = log2(64) = 6

==========

Question 4: **[2 marks]**

Given two binary word vectors w_1 and w_2 as follows:

w_1 = [1010011010]
w_2 = [0011111100]
Compute the Dice and Jaccard similarity between them.

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: **[2 marks]**

Consider two probability distributions for two words be *p* and *q*. Compute their similarity scores with KL-divergence.
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: **[2 marks]**

Consider the following word co-occurrence matrix given below. Compute the cosine similarity between
(i) w1 and w2, and (ii) w1 and w3.

	w4	w5	w6
w1	2	8	5
w2	4	9	7
w3	1	2	3

a. 0.773, 0.412
b. 0.881, 0.764
c. 0.987, 0.914
d. 0.897, 0.315

Answer: c
Solution:

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

Cosine-sim (w1, w2) = (2*4 + 8*9 + 5*7) / (√(2*2 + 8*8 + 5*5) * √(4*4 + 9*9 + 7*7)) = 0.987
Cosine-sim (w1, w3) = (2*1 + 8*2 + 5*3) / (√(2*2 + 8*8 + 5*5) * √(1*1 + 2*2 + 3*3)) = 0.914

==========

Question 7: **[1 marks]**

Which of the following types of relations can be captured by word2vec (CBOW or Skipgram)?

1. Analogy (A:B::C:?)
2. Antonymy
3. Polysemy
4. All of the above

Answer: 1
Solution: Word vectors learnt using CBOW or Skipgram models can’t disambiguate between Antonyms or Polysemous words.

==========