Natural Language Processing

Assignment- 2

TYPE OF QUESTION: MCQ

Number of questions: 10 **Total mark: 10 X 1 = 10**

QUESTION 1:

According to Zipf's law which statement(s) is/are correct?

(i) A small number of words occur with high frequency.

(ii) A large number of words occur with low frequency.

a. Both (i) and (ii) are correct

b. Only (ii) is correct

c. Only (i) is correct

d. Neither (i) nor (ii) is correct

Correct Answer: a

Solution:

QUESTION 2:

Consider the following corpus C_1 of 4 sentences. What is the total count of unique bi-grams for which the likelihood will be estimated? Assume we do not perform any pre-processing.

today is Sneha's birthday
she likes ice cream
she is also fond of cream cake
we will celebrate her birthday with ice cream cake

a. 24
b. 28
c. 27
d. 23

Correct Answer: a

Detailed Solution:

Unique bi-grams are:

<s> today	today is	is Sneha’s	Sneha’s birthday	birthday <\s>
<s> she	she likes	likes ice	ice cream	cream <\s>
She is	is also	also fond	fond of	of cream
cake <\s>	<s> we	we will	will celebrate	celebrate her
her birthday	birthday with	with ice	cream cake	

QUESTION 3:

A 3-gram model is a ___________ order Markov Model.

a. Two
b. Five
c. Four
d. Three

Correct Answer: a

Detailed Solution:

QUESTION 4:

Which of these is/are - valid Markov assumption?

a. The probability of a word depends only on the current word.

b. The probability of a word depends only on the previous word.

c. The probability of a word depends only on the next word.

d. The probability of a word depends only on the current and the previous word.

Correct Answer: a, c, d

Solution:

QUESTION 5:

For the string **‘mash’**, identify which of the following set of strings have a Levenshtein distance of 1.

a. smash, mas, lash, mushy, hash
b. bash, stash, lush, flash, dash
c. smash, mas, lash, mush, ash
d. None of the above

Correct Answer: c

Detailed Solution:

QUESTION 6:

Assume that we modify the costs incurred for operations in calculating Levenshtein distance, such that both the insertion and deletion operations incur a cost of 1 each, while substitution incurs a cost of 2. Now, for the string **'lash'** which of the following set of strings will have an edit distance of 1?

a. ash, slash, clash, flush
b. flash, stash, lush, blush,
c. slash, last, bash, ash
d. None of the above

Correct Answer: d

Detailed Solution:

QUESTION 7:

Given a corpus C_2, the Maximum Likelihood Estimation (MLE) for the bigram "dried berries" is 0.4 and the count of occurrence of the word "dried" is 680. for the same corpus $C_{2,}$ the likelihood of "dried berries" after applying add-one smoothing is 0.05. What is the vocabulary size of C_2?

a. 4780
b. 3795
c. 4955
d. 3995

Correct Answer: a

Detailed Solution:

$$P_{MLE}(berries \mid dried) = \frac{C(dried, berries)}{C(dried)}$$

0.4 = C(dried, berries) / 680
C(dried, berries) = 680*0.4 = 272

$$P_{Add-1}(berries|\ dried) = \frac{C(dried,\ berries) + 1}{C(dried) +\ V}$$

0.05 = (272+1) / (680+V)
V=4780

For Question 8 to 10, consider the following corpus C_3 of 3 sentences.

there is a big garden
children play in a garden
they play inside beautiful garden

QUESTION 8:

Calculate **P(they play in a big garden)** assuming a bi-gram language model.

a. 1/8
b. 1/12
c. 1/24
d. None of the above

Correct Answer: b

Detailed Solution:

P(they | <s>) = 1/3
P(play | they) = 1/1
P(in | play) = 1/2
P(a | in) = 1/1
P(big | a) = 1/2
P(garden | big) = 1/1
P(<\s>|garden) = 3/3
P(they play in a big garden) = 1/3 x 1/1 x 1/2 x 1/1 x 1/2 x 1/1 x 3/3 = 1/12

QUESTION 9:

Considering the same model as in Question 7, calculate the perplexity of **<s> they play in a big garden <\s>.**

a. 2.289
b. 1.426
c. 1.574

d. 2.178

Correct Answer: b

Detailed Solution:

$perplexity = \sqrt[7]{12} = 1.426$

QUESTION 10:

Assume that you are using a bi-gram language model with add one smoothing. Calculate **P(they play in a beautiful garden).**

a. 4.472 x 10^-6
b. 2.236 x 10^-6
c. 3.135 x 10^-6
d. None of the above

Correct Answer: b

Detailed Solution:

|V|=11
P(they | <s>) = (1+1)/(3+11)
P(play | they) = (1+1)/(1+11)
P(in | play) = (1+1)/(2+11)
P(a | in) = (1+1)/(1+11)
P(beautiful | a) = (0+1)/(2+11)
P(garden | beautiful) = (1+1)/(1+11)
P(<\s>|garden) = (3+1)/(3+11)
P(they play in a beautiful garden) = 2/14 x 2/12 x 2/13 x 2/12 x 1/13 x 2/12 x 4/14
= 2.236 x 10^-6

***********END*******