Natural Language Processing

Assignment 2

Type of Question: MCQ

Number of Questions: 10 **Total Marks: 10×1=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

Answer: a
Solution:

Question 2: What is Markov assumption?

i. The probability of a word depends only on the current word.
ii. The probability of a word depends only on the previous word.
iii. The probability of a word depends only on the next word.
iv. The probability of a word depends only on the current and the previous word.

a. i
b. ii
c. iii
d. iv

Answer: b
Solution:

Question 3: 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, calculate the distance between the strings “reading” and “writing”.

a. 8
b. 7
c. 6
d. 5

Answer: c
Solution:

	#	w	r	i	t	i	n	g
#	0	1	2	3	4	5	6	7
r	1	2	1	2	3	4	5	6
e	2	3	2	3	4	5	6	7
a	3	4	3	4	5	6	7	8
d	4	5	4	5	6	7	8	9
i	5	6	5	4	5	6	7	8
n	6	7	6	5	6	7	6	7
g	7	8	7	6	7	8	7	6

The Levenshtein distance is 6

Question 4: How many insertions, deletions and substitutions were required corresponding to the optimal alignment obtained between the strings “reading” and “writing” from Question 3?

a. Insertion: 1, Deletion: 1, Substitution: 2
b. Insertion: 2, Deletion: 2, Substitution: 1
c. Insertion: 4, Deletion: 4, Substitution: 1
d. Insertion: 3, Deletion: 3, Substitution: 0

Answer: a
Solution:

	r	e	a	d	i	n	g
w	r		i	t	i	n	g

Question 5: Consider the following corpus C_1 of three 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. (Consider the beginning of sentence and end of sentence tokens, i.e., <s> and </s>.)

- Julia is visiting the museum
- Julia , Grover and Natasha are friends
- Zoe and Natasha will meet Julia in the museum

a. 23
b. 20
c. 16
d. 18

Answer: b
Solution: The unique bi-grams are :

<s> Julia	Julia is	is visiting	visiting the	the museum
museum </s>	Julia ,	, Grover	Grover and	and Natasha
Natasha are	are friends	friends </s>	<s>Zoe	Zoe and
Natasha will	will meet	meet Julia	Julia in	in the

Question 6: Given a corpus C_2, the Maximum Likelihood Estimation (MLE) for the bigram "computational linguistics" is 0.25 and the count of occurrence of the word "computational" is 1200. If the vocabulary size is 4400, what is the likelihood of "computational linguistics" after applying add-2 smoothing?

a. 0.0538
b. 0.0686
c. 0.0302
d. 0.0444

Answer: c
Solution:
P_{MLE} (linguistics | computational) = 0.25
C (computational) = 1200
| V | = 4400
C (computational , linguistics) = 0.25 $\times$ 1200 = 300

$$\begin{aligned} P_{Add-k}(\text{ linguistics } \mid \text{ computational }) &= \frac{C(\text{ computational , linguistics}) + k}{C\ (\text{computational}) + kV}, k = 2 \\ &\approx 0.0302 \end{aligned}$$

For Question 7 to 10, consider the following corpus C_3 of four sentences.

- **<s> three friends amar akbar and anthony are reading book </s>**
- **<s> amar is reading malgudi days </s>**
- **<s> akbar is reading a detective book </s>**
- **<s> anthony is reading a book by rk narayan </s>**

Question 7: Assume a bi-gram language model. Calculate **P(<s> amar is reading a book </s>)**.

a. 0.0561
b. 0.0625
c. 0.0208
d. None of the above

Answer: c
Solution:
P(amar | <s>) = 1/4
P(is | amar) = 1/2
P(reading | is) = 3/3
P(a | reading) = 2/4
P(book | a) = 1/2
P(</s> | book) = 2/3
P(<s> amar is reading a book </s>) = $\frac{1}{4} \times \frac{1}{2} \times \frac{3}{3} \times \frac{2}{4} \times \frac{1}{2} \times \frac{2}{3} = \frac{1}{48} \approx 0.0208$

Question 8: Given the same language model as in Question 7, what is the Perplexity of the sentence <s> **amar is reading a book** </s> ?

a. 1.804
b. 1.739
c. 1.587
d. 1.648

Answer: b
Solution:
Perplexity = $\sqrt[7]{48}$, when N = number of words.

Question 9: Consider the same Bi-gram model as in Question 7, this time with Laplace (Add-one) smoothing. Calculate **P(<s> akash is reading story book </s>)**.(Consider the beginning of sentence and end of sentence tokens, i.e., <s> and </s> in your vocabulary.)

a. 1.764×10^{-3}
b. 2.242×10^{-7}
c. 0.03
d. None of the above

Answer: d
Solution:
|V| = 19 (<s> and </s> are considered in the vocabulary)
P(akash | <s>) = $\frac{0+1}{4+19}$
P(is | akash) = $\frac{0+1}{0+19}$
P(reading | is) = $\frac{3+1}{3+19}$
P(story | reading) = $\frac{0+1}{4+19}$
P(book | story) = $\frac{0+1}{0+19}$
P(</s> | book) = $\frac{2+1}{3+19}$
P(<s> akash is reading story book </s>) = $\frac{1}{23} \times \frac{1}{19} \times \frac{4}{22} \times \frac{1}{23} \times \frac{1}{19} \times \frac{3}{22} \approx 1.298 \times 10^{-7}$

Question 10: Which of the following sentences has the highest probability estimate assuming Bi-gram model (without smoothing)?

a. S1: <s> anthony loves reading book </s>
b. S2: <s> three friends are reading books </s>
c. S3: <s> akbar and anthony are reading malgudi days </s>
d. S4: <s> amar is reading a detective book by agatha christie </s>

Answer: c
Solution:
P(S1) = P(S2) = P(S4) = 0