Introduction to Large Language Models
Assignment- 2

Number of questions: 8 **Total mark: 6 X 1 + 2 X 2 = 10**

QUESTION 1:
A 5-gram model is a _____ order Markov Model.

a. Constant
b. Five
c. Six
d. Four

Correct Answer: d
Solution: An N-gram model considers only the preceding N −1 words.
An N-gram Language Model ≡ (N −1) order Markov Model

QUESTION 2:
For a given corpus, the count of occurrence of the unigram “stay” is 300. If the Maximum Likelihood Estimation (MLE) for the bigram “stay curious” is 0.4, what is the count of occurrence of the bigram?

a. 123
b. 300
c. 273
d. 120

Correct Answer: d
Solution:

P_{MLE}(curious | stay) = C(stay, curious) / C(stay)
0.4 = C(stay, curious) / 300
C(stay, curious) = 0.4 x 300 = 120

QUESTION 3:

Which of the following are governing principles for Probabilistic Language Models?
a. Chain Rule of Probability
b. Markov Assumption
c. Fourier Transform
d. Gradient Descent

Correct Answer: a,b
Solution: Probabilistic Language Models exploit the Chain Rule of Probability and Markov Assumption to build a probability distribution over sequences of

words.

For Question 4 to 5, consider the following corpus:

<s> the sunset is nice </s>
<s> people watch the sunset </s>
<s> they enjoy the beautiful sunset </s>

QUESTION 4:
Assuming a bi-gram language model, calculate the probability of the sentence:
<s> people watch the beautiful sunset </s>

Ignore the unigram probability of P(<s>) in your calculation.

a. 2/27
b. 1/27
c. 2/9
d. 1/6

Correct Answer: a
Solution:

*P(<s> people watch the beautiful sunset </s>) = P(<s>) * P(people | <s>) * P(watch | people) * P(the | watch) * P(beautiful | the) * P(sunset | beautiful) * P(</s> | sunset)*

Ignoring the leading unigram probability of P(<s>), we have:

*P(<s> people watch the beautiful sunset </s>) = P(people | <s>) * P(watch | people) * P(the | watch) * P(beautiful | the) * P(sunset | beautiful) * P(</s> | sunset)*

The conditional probability P(y | x) is calculated according its MLE as:
P(y | x) = Count(x, y) / Count(x)

P(people | <s>) = 1/3
P(watch | people) = 1/1
P(the | watch) = 1/1
P(beautiful | the) = 1/3
P(sunset | beautiful) = 1/1
P(</s> | sunset) = 2/3

Thus, P(<s> people watch the beautiful sunset </s>) = ⅓ x 1 x 1 x ⅓ x 1 x ⅔ = 2/27

QUESTION 5:
Assuming a bi-gram language model, calculate the perplexity of the sentence:
<s> people watch the beautiful sunset </s>

Please do not consider <s> and </s> as words of the sentence.

a. $27^{1/4}$
b. $27^{1/5}$
c. $9^{1/6}$
d. $\left(\frac{27}{2}\right)^{\frac{1}{5}}$

Correct Answer: d
Solution:

As calculated in the previous question,
P(<s> people watch the beautiful sunset </s>) = $\frac{2}{27}$
Ignoring <s> and </s>, total number of words in the sentence = 5
Thus, Perplexity = $\left(\frac{27}{2}\right)^{\frac{1}{5}}$

QUESTION 6:

What is the main intuition behind Kneser-Ney smoothing?
a. Assign higher probability to frequent words.
b. Use continuation probability to better model words appearing in a novel context.
c. Normalize probabilities by word length.
d. Minimize perplexity for unseen words.

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

QUESTION 7:

In perplexity-based evaluation of a language model, what does a lower perplexity score indicate?
a. Worse model performance
b. Better language model performance
c. Increased vocabulary size
d. More sparse data

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

QUESTION 8:

Which of the following is a limitation of statistical language models like n-grams?

a. Fixed context size
b. High memory requirements for large vocabularies
c. Difficulty in generalizing to unseen data
d. All of the above

Correct Answer: d
Solution: N-gram models suffer from fixed context size, data sparsity, high memory usage, and inability to generalize well to unseen data.
