Language modeling [Solution by Karthikeyan.S]

LATEST SUBMISSION GRADE

100%

1.Question 1

Given the corpus of three sentences

This is the house that Jack built.

This is the malt that lay in the house that Jack built.

This is the rat that ate the malt that lay in the house that Jack built.

calculate the probability *p* (lay | that) using maximum likelihood estimation.

●

1/3

○

1/2

○

3

○

2/3

Correct

Exactly! There are six occurrences of "that", and only two of them are followed by "lay".

1 / 1 point

2.Question 2

Consider the **bigram language model** trained on the sentence:

This is the cow with the crumpled horn that tossed the dog that worried the cat that killed the rat that ate the malt that lay in the house that Jack built.

Find the **probability of the sentence:**

This is the rat that worried the dog that Jack built.

○

1/8

○

\infty∞

○

0

○

\frac1{2} \cdot \frac1{3} \cdot \frac1{6} \cdot \frac1{2} \cdot \frac1{7} \cdot \frac1{2} \cdot \frac1{6} \cdot \frac1{2} \cdot \frac1{7} \cdot \frac1{3} \cdot \frac1{5} \cdot \frac1{4}21·31·61·21·71·21·61·21·71·31·51·41

●

\frac1{6} \cdot \frac1{7} \cdot \frac1{6} \cdot \frac1{7}61·71·61·71

Correct

Exactly! Most of the conditional probabilities are equal to 1, e.g. p(is|This) = 1 since "This" occurs only once in the training data and it's followed by "is". Only the probabilities for "the" and "that" are non-trivial.

2 / 2 points

3.Question 3

Consider the **trigram language model** trained on the sentence:

This is the cat that killed the rat that ate the malt that lay in the house that Jack built.

Find the **perplexity** of this model on the test sentence:

This is the house that Jack built.

○

\frac{1}{\sqrt[7] {\frac{1}{3} \cdot \frac{1}{3}}} = \sqrt[7] {9}731·311=79

○

1

○

0

●

∞

Correct

Yes. The probability *p* (house | is the) is zero.

1 / 1 point

4.Question 4

Apply **add-one smoothing** to the trigram language model trained on the sentence:

This is the rat that ate the malt that lay in the house that Jack built.

Find the **perplexity** of this smoothed model on the test sentence:

This is the house that Jack built.

Write the answer with precision of 3 digits after the decimal point.

10.205

Correct

Exactly! You did a good job!

4 / 4 points

5.Question 5

Find one incorrect statement below:

●

If a test corpus does not have out-of-vocabulary words, smoothing is not needed.

○

Trigram language models can have a larger perplexity than bigram language models.

○

N-gram language models cannot capture distant contexts.

○

End-of-sentence tokens are necessary for modelling probabilities of sentences of different lengths.

○

The smaller holdout perplexity is - the better the model.

Correct

Even though the probabilities will not be equal to 0, they will be still poorly evaluated for rare terms!