Language modeling

LATEST SUBMISSION GRADE

60%

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 (house | the) using maximum likelihood estimation.

- 2
- 1/6
- 1/2
- 1/3

Correct

This is correct. There are six occurrences "the" in the corpus and only three of them are followed by "house".

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.

- 0

- ∞
- 1/2 * 1/3 * 1/6 * ½ * 1/7 * ½ * 1/6 * ½ * 1/7 * 1/3 * 1/5 * 1/4
- 1/8
- 1/6 * 1/7 * 1/6 * 1/7

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 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.

- 0
- ∞
- 7th root of 9
- 1

Correct

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

1 / 1 point

5.Question 5
Find one incorrect statement below:

- If a test corpus does not have out-of-vocabulary words, smoothing is not needed.
- N-gram language models cannot capture distant contexts.
- The smaller holdout perplexity is - the better the model.

- Trigram language models can have a larger perplexity than bigram language models.
- End-of-sentence tokens are necessary for modelling probabilities of sentences of different lengths.

Correct

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