

Out-of-vocabulary words

Toy train corpus:

This is the house that Jack built.

Toy test corpus:

This is the *malt*.

What's the perplexity of the Bigram LM?

$$p(\text{malt}|\text{the}) = \frac{c(\text{the malt})}{c(\text{the})} = 0$$

$$p(\mathbf{w}_{\text{test}}) = 0$$