Leo Goldstein

1. Define N gram, examples for unigram, bigram, trigram
    a. N-gram is a contiguous sequence of N words in a text.
    b. Examples for unigram, bigram, trigram
        i. Unigram: [‘I’, ‘love’, ‘natural’, ‘language’, ‘processing’]
        ii. Bigram: [(‘I’, ‘love’), (‘love’, ‘natural’), (‘natural’, ‘language’), (‘language’, ‘processing’)]
        iii. Trigram: [('I', 'love', 'natural'), ('love', 'natural', 'language'), ('natural', 'language', 'processing')]
    c. Concept of BOS, EOS, UNK
        i. BOS is Beginning of Sentence. Ex: [‘<BOS>’, ‘I’, ‘love’, ‘NLP’]
        ii. EOS is End of Sentence. Ex. [‘I’, ‘love’, ‘NLP’, ’EOS>’]
        iii. UNK is unknown token, words not in vocabulary. Ex: [‘I’, ‘love’, ‘<UNK>’] where NLP is not in vocab
2. Probability and smoothing
    a. Count(“cat sat”) = 1
    Count(“cat”) = 1
    MLE Formula: Count(“cat sat”)/Count(“cat”) = 1/1 = 1.0
    b. Smoothing prevents zero probabilities when an unseen N-gram appears, ensuring the model can generalize better
    Laplace smoothing: adds 1 to all counts to avoid zero probabilities
    c. P(“sat” | “cat”) = (Count(“cat sat”) + 1)/(Count(“cat”)+V) = (1+1)/(1+7) = 0.25
3. Intrinsic and Extrinsic Evaluation
    a. Intrinsic evaluation: tests a model on an isolated task
    Extrinsic evaluation: tests a model’s effectiveness in a real world task
    b. Perplexity measures how well a model predicts unseen text. Lower perplexity = better language model
    c. PP = (0.01)^(-1/10) = 10^0.1 = 1.2589