CSE 143 HW2

Maximum Likelihood Estimate

Arnav Nepal, Harsh Jha

May 15, 2024

1 N-gram models

1.1 Unigram Model: WIP

Our WIP unigram model is implemented in the FeatureExtractor class and it's children in utils.py. As of now, we only have the Unigram model properly tested, but there is code for the Bigram and Trigram models. In our `UnigramFeature` class, the `fit()` function is used to associate tokens with a specific index in the `self.unigram` dictionary. Afterwords, we tranform the dictionary using the transformation functions to obtain a feature array. The feature array is then fed into the `fit()` function in the `MaximumLikelihoodEstimateUnigram` class in classifiers.py to obtain an array of the probabilities for each type in the data set. At this point, we are ready to test on data, and perplexity scores can be obtained using the `perplexity()` function in the `MaximumLikelihoodEstimateUnigram` class.

Currently, the our program solely prints the Unigram perplexity score on the debug text provided in the assignment specification (`HDTV . <STOP>`).

2 Organization

We organized our code by making separate classes for the Unigram, Bigram, and Trigram models. These models are in `utils.py`. We also made a separate class to calculate the Maximum Likelihood Estimate for each of the models. The MLE classes are in `classifiers.py`. The `main.py` file contains the main code for running our models and printing out pertinent information to I/O.