Repeat Assignment – Natural Language Processing

Data

The training data consist of 100 sentences and the test data of 10 sentences, both generated by the same probabilistic grammar:

- Training data
- Test data

Questions

[50 marks]

1. Check what probability the different models assign to the sentences in the test set. You should test the unigram, bigram and trigram models and compare them to each other and to the true probabilities. Go through all ten sentences with all three models and study how the probabilities vary for different models and sentences. Check how they relate to the true probabilities for each sentence and try to explain the patterns you find. (Python program should be attached separately)

[50 marks]

2. The normal way of evaluating a language model is to compute the probability or entropy (perplexity) that the model assigns to a given test corpus. The principle is that a better model gives higher probability (lower entropy). To compute these metrics, by replacing unigram by bigram and trigram, you can test all three models. Compare the result for a given model with the true probability (entropy) given for the test data. (Python program should be attached separately)