**CSE 446 – Machine Learning – Project 3.2**

**Original Accuracies**

* Mini Dataset
  + Training Accuracy: 98.8%
  + Test Accuracy: 50.68%
* Large Dataset
  + Training Accuracy: 93.82%
  + Test Accuracy: 81.65%

**POS Tagger Accuracies**

* Mini Dataset
  + Training Accuracy: 98.6%
  + Test Accuracy: 72.55%

**Improvement #1 – Remove most common words**

* Mini Dataset – Optimized at removal of 500 most common words
  + Training Accuracy: 99.5%
  + Test Accuracy: 65.59%
* Large Dataset – Optimized at removal of 800 most common words
  + Training Accuracy: 95.37%
  + Test Accuracy: 84%

**Improvement #2 – Remove least common words**

* Mini Dataset – Optimized at removal of 7000 least common words
  + Training Accuracy: 98.5%
  + Test Accuracy: 51.27%
* Large Dataset – Optimized at removal of 7000 least common words
  + Training Accuracy: 93.82%
  + Test Accuracy: 81.8%

**Improvement #3 – Transformation 4.2 – By Document Frequency**

* Mini Dataset
  + Training Accuracy: 99.6%
  + Test Accuracy: 57.65%
* Large Dataset
  + Training Accuracy: 95.73%
  + Test Accuracy: 83.3%