Language Engineering Assignment 2
Meyer Dauber - Spring 2020

Problem 3a.

Model	Test Corpus	# of Words	Estimated Entropy
Kafka	Small	19	10.29
Small	Kafka	24944	13.46

Problem 3b.

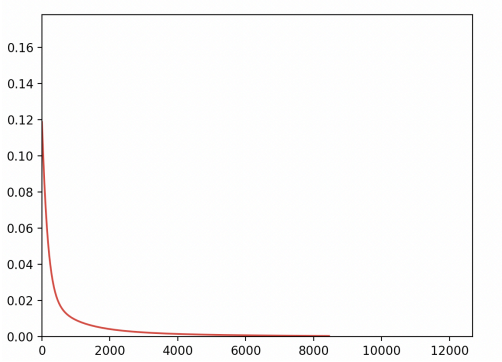
Model	Test Corpus	# of Words	Estimated Entropy
Austen	Austen	13161	5.72
Austen	Guardian	871837	9.75
Guardian	Austen	13161	6.40
Guardian	Guardian	871837	6.62

Conclusions:
- Might be harder to predict Guardian documents because the articles are written by different people with different styles whereas Austen is homogeneous
- Guardian model has less variability in entropy across test materials, potentially because the number of unique bigrams is higher

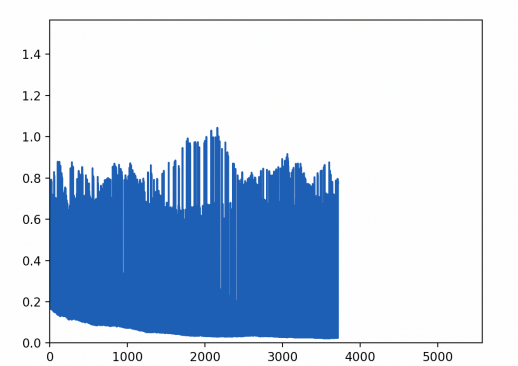
Problem 4b.

Test Set Results: (LR = 1)

Batch
Theta: [-2.282728016, 6.989215584, -5.372798763]
Gradients: [0.000019311, -0.000992743, 0.000999297]

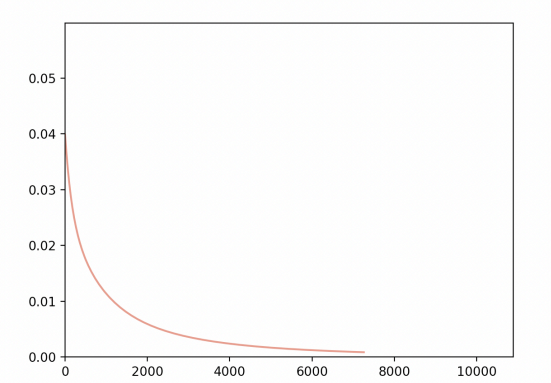


Stochastic
Theta: [-1.095996640, 8.939071303, -5.185648300]
Gradients: [-0.000392615, -0.000392615, -0.000000000]



Performs faster than Batch, but theta values vary wildly, meaning accuracy is likely low

MiniBatch
Theta: [-2.282720755, 6.988842223, -5.372422937]
Gradients: [0.000019325, -0.000993102, 0.000999661]



Minibatch converges much faster and more accurately than the other 2 methods of GD

Accuracy = (True Positive + True Negative)/(True Positive + False Positive + True Negative + False Negative)

Precision = (True Positive)/(True Positive + False Positive)

Recall = (True Positive)/(True Positive + False Negative)

Batch @ LR = 0.01
-2.282694747 6.987514073 -5.371086145
0.000019348 -0.000993433 0.000999998
Runtime: 28.806554794311523 seconds

Stochastic
-2.165150415 9.087358563 -7.269657743
-0.000984700 -0.000984700 -0.000000000
Runtime: 15.10145616531372 seconds

MiniBatch
-2.282694763 6.987514897 -5.371086974
0.000019348 -0.000993432 0.000999998
Runtime: 30.401247262954712 seconds