ESE 545 Project 2

Andrew Hunter, John Waldt

Question 1.1. The csv of training data was read in as a pandas dataframe and the Sentiment labels were changed to −1 and 1. □

Question 1.2. The cleaning process only included steps 1-6 and the given stop word list was used. All tweet strings were cleaned using regular expressions and pandas replace method for dataframes. Then each cleaned tweet string was split into a list of words by spaces. □

Question 1.3. Unigram and Bigram features were created using CountVectorizer from the sklearn package. For unigram case, our training set was $n = 1600000$ and $p = 383587$. For bigram, we had $n = 1600000$ and $p = 3395410$. We scaled all features by their standard deviations. □

Question 1.4. See Figure 1. □

Question 1.5. See Figure 1. □

Question 1.6. The test set was loaded in and cleaned the same way the training set was. It was then turned into a bag of words using the vocabulary from the training set.

Unigram: Due to the relatively small sample size of the test set ($n = 498$), some randomness can be seen in the test error. In the end we achieved around a 20% error on both the test and training set with Pegasos and AdaGrad. For both Pegasos and AdaGrad, the inputs were $\lambda = .009$, Batch Size = 4500. Figure 1 shows a plot of unigram error rate by number of iteration and Table lists the final accuracy and error rates.

Bigram: Using the bigram, we were able to achieve a training error of under 10% for both Pegasos and AdaGrad. However, the accuracy on the test set was only slightly better than random guessing. This is most likely due to over-fitting the training sample, and a very sparse feature matrix for the test set.For both Pegasos and AdaGrad, the inputs were $\lambda = .001$ and Batch Size = 4500. Figure 2 shows a plot of bigram error rate by number of iteration and Table lists the final accuracy and error rates. □

Unigram Results		
	AdaGrad	Pegasos
Train Error	18.03%	20.86%
Test Accuracy	85.34%	83.13%

Table 1: Unigram Final Results

Bigram Results		
	AdaGrad	Pegasos
Train Error	5.07%	8.4%
Test Accuracy	56.8%	56.2%

Table 2: Bigram Final Results

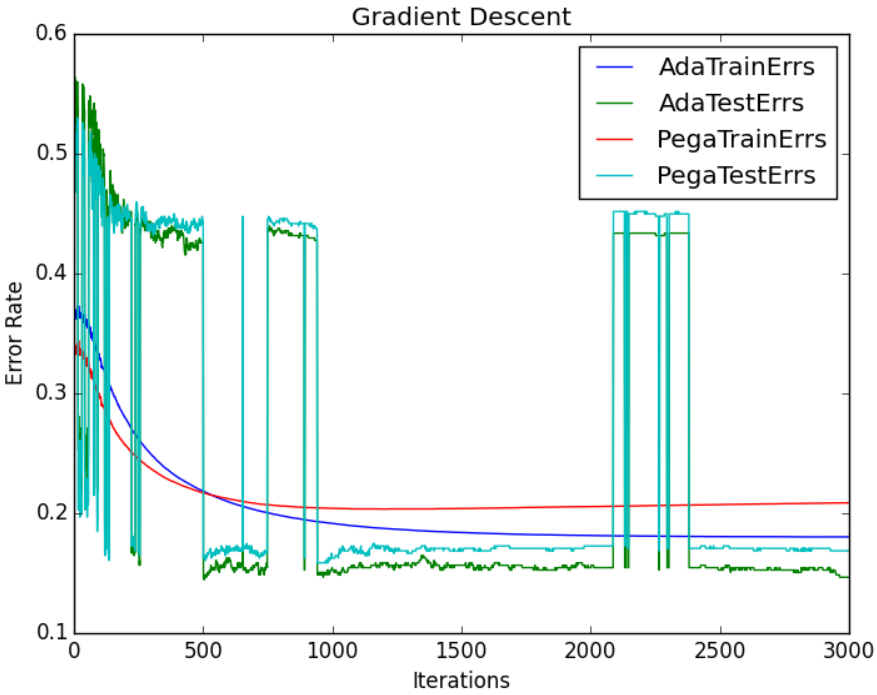

Figure 1: Unigram Plot

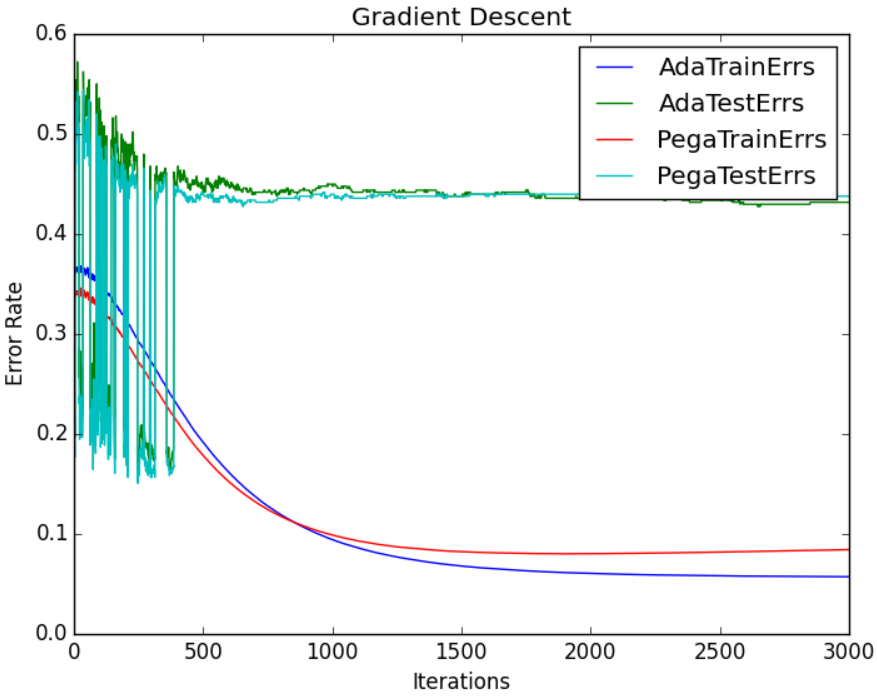

Figure 2: Bigram Plot