11-761 Language and Statistics - Spring 2013
Course Project

Leonid Boytsov, Zhou You, Di Wang, Hector Liu

Abstract

Our experiments indicate that an n-gram-based approach to synthetic documents outperforms other methods. It allows us to achieve the accuracy 0.9 using the provided training set and the slightly higher accuracy 0.91, if we train our model on the larger training data (created by our group). By using the 3-gram and 4-gram models alone, one can get accuracy as high as 0.88. In the case of the soft metric, additional data allows us to improve the average likelihood by a larger margin, namely, from 0.78 to 0.89. We use a generative approach and train different models (the soft-margin SVM and the L1-regularized logistic regression) to produce class labels and likelihoods.

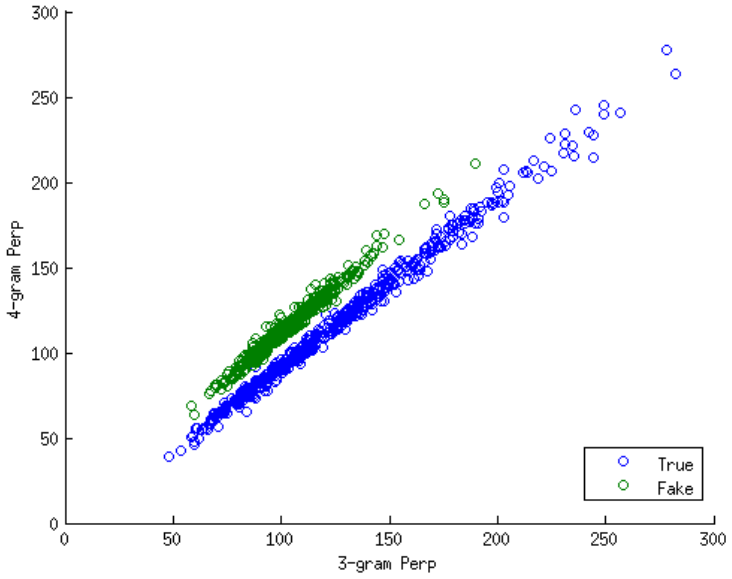

Figure 1: Perplexities for 3- and 4-gram models.