Ethnicity Classifier

This classifier uses the sub-strings of size 3 to learn probability associated with each substring such that it belongs to a particular ethnic group using the census 2000 data. When a new string is presented, it is broken down into sub-strings of size 3, and prediction is done on each of the sub-string using the model computed before. The output accuracy of the predictor is 85%. The model can be improved by balancing the sub-strings and the across the ethnic groups. The model follows the Naive Bayes assumption of conditional independence.