Machine Learning Final Project Progress Report

Michael Ashmead (mashmea1@jhu.edu)
Maxwell Yeo (myeo1@jhu.edu)

November 28th, 2016

So far we have turned both sets of data into instances consistent with the method applied to the datasets we used for our homework. One dataset is anonymized Reddit posting behavior. Each row is a different user that consists of their User ID and then name of the subreddits they commonly post in (all comma-separated). We first created a list of all subreddits that appear in our dataset and assigned each a unique number. Then we converted each comma-separated subreddit into its assigned number and used it as a key in our Feature Vector and set it's respective value to 1. Our second dataset was a list of all subreddits and their publically available information like descriptions, header images, if over 18 only, etc. This data set was 1.71GB large, so the first thing we did was create a new document that just listed the subreddit and its description. Second we created a corpus of words that appeared in the descriptions. We got a list of 124,707 words and assigned each an index. Then we created an instances file that listed each subreddit followed by the word frequency of its description.