Result Analysis for ID1 Batch 1 Full scale

Koushik Kumaraswamy

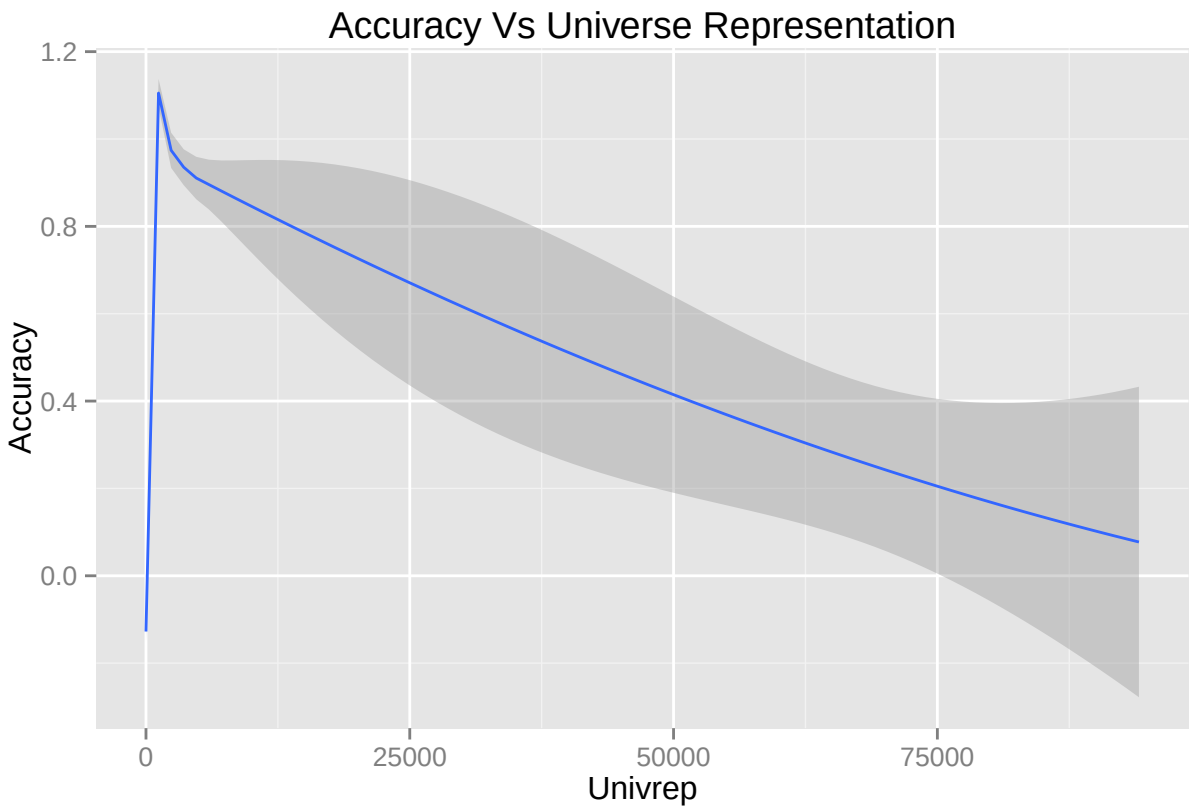
Thurday, Apr 30, 2015

Background

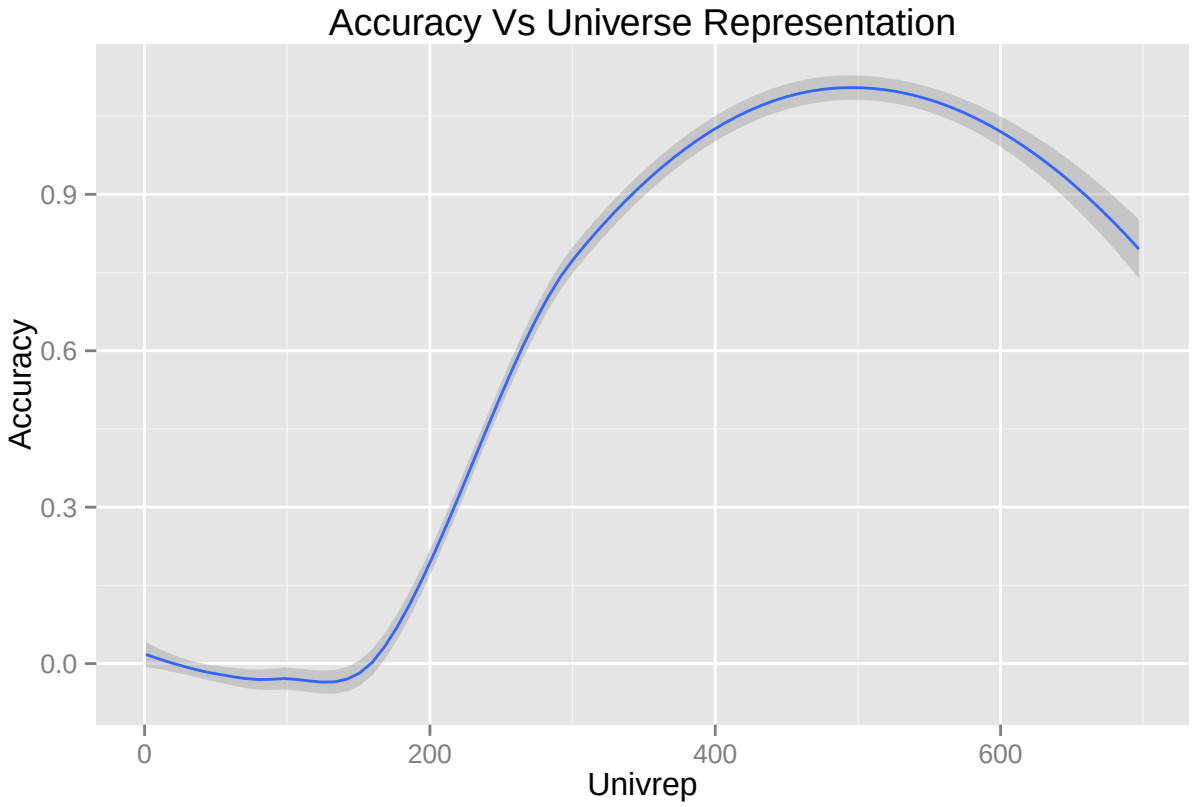
In this report, we evaluate the performance of the Vector Space Model, (referred to as VSM henceforth), at full scale for ID #1, with a view of parsing the relationship betweem model accuracy and Universe representation

Run Results and analysis

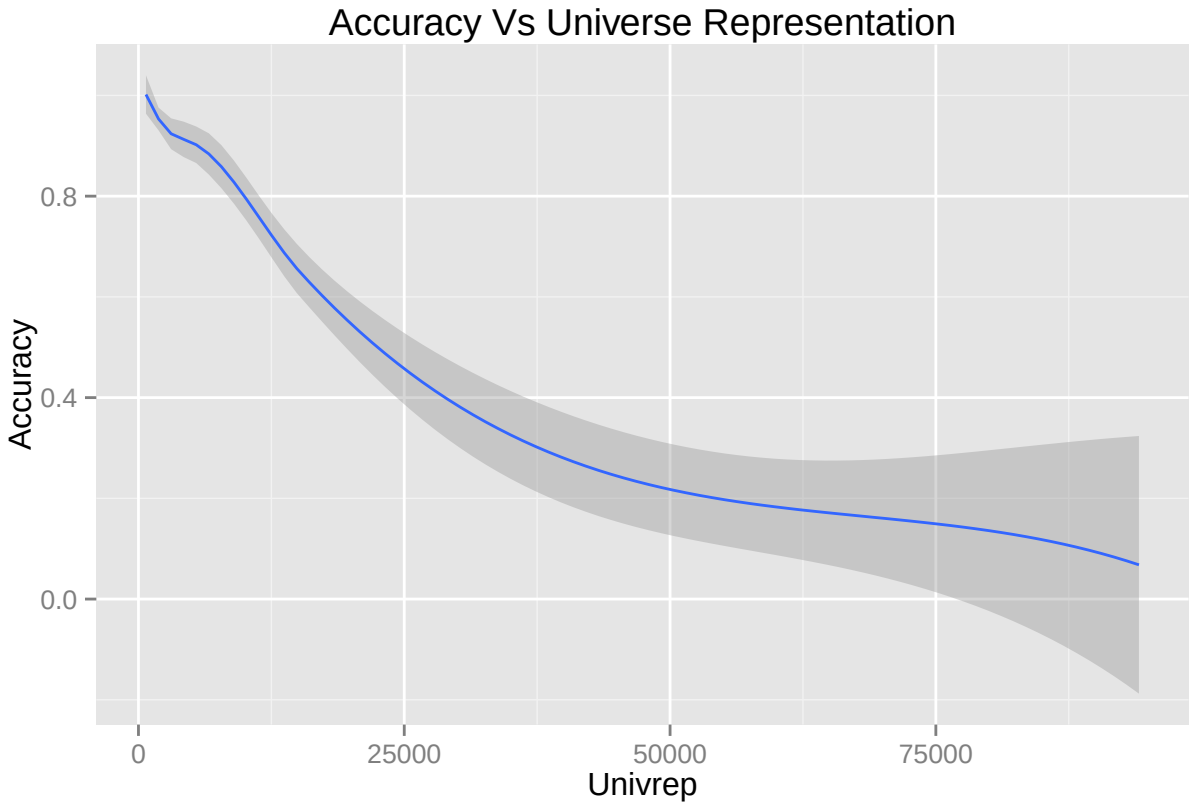
Plotting accuracy as a function of universe representation we get:



From the graph above, there seems to be two seperate trends we can observe.One increasing and one decreasing. We can *cut* the data set at the point of change, and observe the trends seperately as below. For the first part of the graph, we see..



For the second part of the graph, we see the following...



Analysis of data

From the above, and additional offline analysis, we observe that the model accuracy seems to be 95+% if training set count is between 700-1000. When the training set count is between 600-1200, the accuracy falls to 80+%. The accuracy further falls to 65+% if the training count varies between 500-3500. This seems to suggest that there is a sweet spot for the classifier in terms of number of records per group that it seems to do well with. This finding needs to be validated through additional randomized runs.