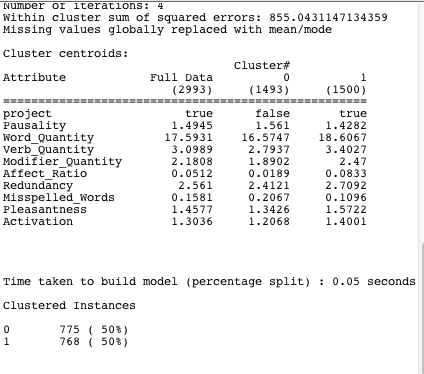
Weka Analysis

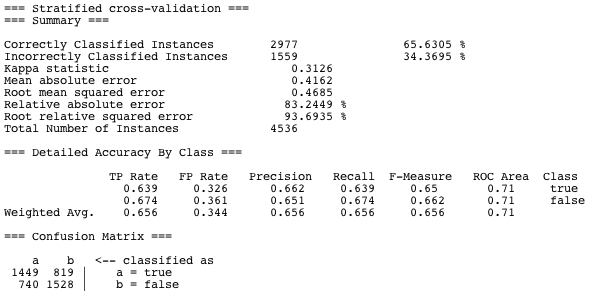
To start this analysis, I loaded the file Nicole had created into Weka and (in the preprocess tab) started by removing attributes I was sure weren't connected with project prediction such as the user and partner ids. Next I removed the attributes where all of the values were zero. Then I used the visualization option on the preprocess tab to visualize each attribute as it related to the project category, and removed attributes with very small values.

I was doing exploratory analysis, so I did a SimpleKMeans Cluster with a 66% split to isolate the attributes with the largest differences between project and non-project data sets. The printout from this analysis is included as figure 1 below. These attributes are Pausality, Word\_Quantity, Verb\_Quantity, Modifier\_Quantity, Affect\_Ratio, Redundancy, Misspelled\_Words, Pleasantness, and Activation. Some of these are more interesting than others, such as Affect\_Ratio and Misspelled\_Words, were particularly interesting to us.

**Figure 1**



At this point, I narrowed the data file to include only the project attribute and the nine other attributes with the biggest difference between project and non-project comments. With the new version of the data file loaded in Weka, I started running various different classifiers. So far the best is LMT with a correct rating of 65.6305%, but with the narrowed field all classifiers are successful ~55-65 % of the time. The output from the LMT analysis is shown below.

**Figure 2** 

This output shows that the LMT analysis was able to successfully classify 65.6305% of the instances (or 2977 of the 4536 comments). The output also includes the accuracy by class, showing the true- and false- positive rates for both the project (true) and non-project (false) comments. The true positive for each class indicates the rate of instances which were accurately classified as belonging to that class (project comments classified as project comments) while the false-positive indicates the rate of instances which were falsely classified as belonging to that class (non-project comments classified as project comments). From this analysis, it appears that the non-project class has higher positive rates for both true- and false- positive classifications. This suggests that the analysis might default to classifying a comment as non-project.