Magellan User Experience and Feedback

Ashish Shenoy (ashenoy), Shruthi Racha (shruthir)

1. **Documentation for the Magellan Library:**
   The developer manual and the user manual needs to be updated. Many of the function signatures and the examples are not up-to-date.

2. **False Negatives and False Positives:**
   Along with the number of false negatives and false positives, the function mg.print_eval_summary(eval_result) should let the user know what the false negative and the false positive tuples are. This will allow us to quickly know if the labelling for those two tuples were wrong.

3. **Handle spaces in the column names of tables:**
   Magellan could not handle spaces in the attribute names. For example, when we first tried to run the attribute equivalence blocker on our tables, the Phone Number column had space in it and the blocker failed without giving any indication that it failed due to the space in the column name. It should also be able to handle trailing and leading spaces in the names of the attributes. Considerable time was spent in figuring out why "Name" and " Name" were being recognized as same.

4. **Exception and Error Handling:**
   This is one area which needs considerable improvement. Cases like wrong usage of the function calls or wrong parameters should fail with an informative error message which would point to the error. Right now we need to skim through a stack trace and need to manually examine the line number in the code to understand what is going wrong.

5. **Debug Blockers:**
   The debug blockers should use the same internal logic as the actual blockers. While trying to understand why a particular blocker was not working as expected when the debug blocker was working fine, we learned that the debug blockers use a different workflow when compared to the actual blockers. This defeats the purpose of debugging the actual blockers.

6. **Labelling the table for golden data:**
   There should be an option to delete a particular row/tuple from the golden data while labelling them. This would be useful in scenarios where some of the tuples might not have been cleaned up properly and we want to get rid of it and not affect the learning.

7. **Metrics while selecting best matcher:**
   While selecting the best matcher, we would have liked to have the ability to see all three metrics : Precision, Recall and F1 score in one shot rather than running mg.select_matcher separately for the three metrics. Maybe mg.select_matcher could

take in display_metrics and selection_metric as parameters, one for displaying the metric values and the other to select the best matcher.

8. **Explanation of the features used train the matchers:**
   Need documentation on the features being used to train the matchers. For example, we couldn't figure out clearly what *NAME_NAME_mel* is measuring.

9. **Function to remove features from the feature vector:**
   A function to remove features from the feature vector using the feature ID or the feature name would be a good thing to have. Currently we had to split the vector and concatenate them using the inbuilt python functions. But abstracting this part in Magellan would have resulted in a better user experience.

10. **The trigger needs a debugger too:**
    To understand what exactly a trigger is doing, a debugger with the tuples that the trigger acted on highlighted would help a developer understand what exactly the trigger is doing if he is not sure about the conjunctions and the disjunctions of the rules added to the trigger.

11. **Magellan does not work on the newer versions of MAC:**
    We struggled to get Magellan running on our Macs. The Magellan library which uses pandas, fails to import pandas. This seems to be a popular bug in the anaconda package but we couldn't figure out how to work around it.