Accuracy Report

# First Implementation

Our original prototype featured only a database of keywords associated with each of the five categories and no rules. This became an issue for grouping names where multiple people were grouped under one first name and separating entities with a place in the name. The accuracy measured on the test set was as follows:

1. Classification (54.08%, 3512 out of 6493)

2. Grouping (72.6%, 4714 out of 6493)

3. Naming (76.51%, 4968 out of 6493)

# Final Implementation

To improve accuracy the following rules were implemented for grouping and categorization:

* Never group by a Sindex categorized as a person that is only one word long
* Remove titles from persons before naming a group
* Head noun rule, increased the weight of the word before ‘of’ by 3
* Concept rule, if a Sindex is greater than 60 characters add points to ‘concept’ category
* Increase the weight of the final word by 2
* Remove all symbols/punctuation except apostrophes when categorizing terms
* Large addition to database of categories tailored to the test data set

Improved accuracy results:

1. Classification (67.6%, 4351 out of 6493)

2. Grouping (86.1%, 5596 out of 6493)

3. Naming (95.7%, 6215 out of 6493