Project proposal

Team Members:

Chukwubuikem Ume-Ugwa

Text Corpus: Corpus

The corpus I plan to use is a corpus that is a collection of reddits threads used in this paper to classify comments in online discussions into a set of coarse discourse acts. I am choosing this corpus because I want to see if I can reproduce the same results as the paper.

Rationale:

The rationale behind the project is to see how well I can replicate the work of others and to improve my understanding of the paper. I plan to implement a system that given a query will classify that query into one of ten discourse acts identified in the paper. This project falls in the text categorization section of the syllabus.

The project will follow the typical methodology for creating a language model. First, I will have to create multiple transformations of the data to find the transformation that yields the best performance. This transformation will be done in conjunction with model selection. This model selection helps in picking the best model for the task before proceeding to the training phase.

Once a model has been selected, the training phase begins. At this point, the data transformation is now fixed. After training the selected model and the performance validated, I will then build an interface through which queries will be given to the model and the output will be the discourse act classification for the query.