Tolkenizer

CS 551 Project Proposal

Akanksha Vyas & Samuel Payson

October 27, 2011

This proposal describes our idea for the CS 551 final project. The goal of the project is to create an AI system that, given words from a particular language, will produce similar words based on the probabilities of different letters following one another.

First we will declare *token* type that could be mono-, bi-, or tri-graphs. From a test data set the program will learn the probabilities of tokens following each other, and use this data to create a hidden Markov model. Eventually, given tokens as inputs the program should be able to create words resembling the learning data set. We will implement this project in the Go Programming Language.

Current implementations of similar projects use trivial methods like scrambling data and then using random number generators. We are hoping that this method could have applications to natural language.

The project is hosted on https://github.com/vyasa/Tolkienizer