Markov Text Generator Writeup

Rocky Xia

General:

The Markov Text Generator is able to read and process a corpus of any size and generate a piece of text similar in style to the training corpus. The size of the corpus would directly influence the output's similarity to the input, as well as its range of vocabulary.

Structure:

The algorithm would consist of three major parts, the CorpusReader, MarkovChain, and TextGenerator. Each of the class objects would be explained in detail in the following:

CorpusReader:

CorpusReader is a class object that reads in a corpus from a text file and stores it at construction, while containing many methods for data processing. These methods include a tokenizer that returns an array of strings contained in the corpus, a method that returns the number of unique words, etc. This would be the method called to initialize the Markov chain before starting text generation.

MarkovChain:

MarkovChain is a class object that constructs a Markov chain data structure from a CorpusReader object. The Markov chain would construct an array of Entity objects, which contain a String that represents a word and a linked list of indices of words that appear after the String in the corpus. The Markov chain would have a method that takes in one parameter, when called, returns a word at random that succeeds the String entered prior.

TextGenerator:

TextGenerator is the class object where users have access to a console interface. They will be given the option to select their own corpus, or choose one that is already loaded in. The program would then use CorpusReader and MarkovChain to create a body of text similar to the training corpus.