Project Members: Andres Orozco

Topic: Applications of Markov Chains

Proposal:

Google's Page Rank algorithm is the most widespread application of Markov chains in the real world. However, the applications of Markov chains extend beyond ranking webpages. Markov Chains "describe a sequence of possible events in which the probability of each event depends only on its state attained from previous events" [1]. So hypothetically, any discrete system which "hops" from state to state based on probabilities can be modeled using Markov Chains.

One discrete system which markov chains can act on is language. Given a set of words/sentences, we can calculate the probabilities that one words leads to another. Example:

- The → dog (0.5)
- The → cat (0.5)

Given the word "the" we know that the next word must be either dog or cat. If we extend this idea to cover a longer set, say a tweet or collection of tweets, it is possible to create markov generated tweets based on the aforementioned probabilities between words. This can obviously be extended to longer sets, such as paragraphs or books.

My project proposal is to see whether markov chains can model real human language to the point where it is indistinguishable from the true data set. Goals for this project:

1. Does the output remotely resemble human language?
2. If so, would someone else have trouble picking the "right" option?

Markov Chains require a data set, that is a set of probabilities from which to construct the transition matrices. In the page rank example, the probabilities of links to each page is what constitutes

1 Definition of Markov Chains

the matrix. For language, the probability that a given word leads to another would constitute the transition matrix. Some of the data sets I have been thinking of generating include:

- Wine reviews
- Tweets from unique individuals
- Poetry
- Headlines
    - Compare and contrast different news outlets' markov outputs.