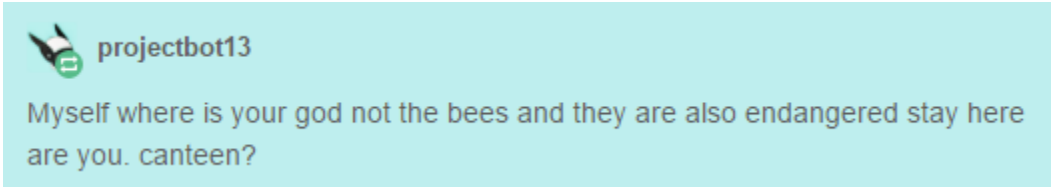
Proposal: Twitter Sentence Generator

Derek Chaplin ———— COMP 3770-01

For my artificial intelligence project, I will attempt to create a program which runs a Twitter account that will post its own Tweets. This project will mainly deal with natural language processing (NLP), which can either be somewhat easy to implement or incredibly difficult depending on how decent you want the sentences to be. It will implement Markov chains using existing works of writing in the mystery genre to generate sentences.

NLP is a very important field in computer science because while creating programs and robots to do things automatically can be nice, it'd be nicer if we can make them communicate with us. In fact, human communication is the most important part of passing the Turing Test.

While bots that write stories, social media posts, and even poetry, on sites like Tumblr do exist, they usually spit out unintelligible sentences. Here's an example:



These bots are usually towards the "messy but it works" end of the spectrum of NLP. Most typically use what's known as Markov Chains, a stochastic model, to structure their sentences. They start their sentence with a word and then randomly pick the next word based on associated words. However usually these implementations don't use existing works to influence their writing. They typically keep their own database of words and associated words and try to guess a sentence structure randomly. The random generation with this method produces very poor, most times grammatically illegal, sentences.

The method I will attempt to use goes further than how these bots are usually structured. I'll gather up a bunch of existing works from the mystery genre (to give my bot a certain style) and have my program process these works to mimic the style of writing in its own tweets using Markov chains. It will use these existing works to determine the next word in the sentence, which I believe will produce more coherent results if given a big enough amount of existing data.

First, I need to collect as much data for my program to mimic in the mystery genre and to process and get it in a consistent format. I'll also add commonly used names for characters and settings and such. This will probably take up to a week depending on how much data I end up using. Then, I'll write my own implementation of Markov chains make my program process these works and determine the probability of the next word it should pick when given a starting word (about two weeks). I'll then create the Twitter bot account and integrate the Twitter API into my code and set it to run and produce constant example results. I can then tweak my code based on results to generate better results. (about two weeks).

I intend to work on this project by myself.