Marion Madanguit
Software Design
March 9, 2020

Text Mining and Analysis Reflection

Project Overview

For this mini project, I performed Markov analysis on 50 movie scripts from the Internet Movie Script Database (IMSDb) to generate my own 500-word movie script. By reading through several of these computer-generated movie scripts, I was also able to study the limitations of a simple Markov model.

Implementation

My project consisted of two main tasks: 1) harvesting scripts from the internet and 2) analyzing those scripts using Markov analysis. I chose to harvest scripts from the Internet Movie Script Database (IMSDb) because of the wide range of movie scripts it has available and also the simplicity of its formatting. Each script was presented plainly on each page, which made its HTML especially easy to parse.

For this project, I decided to generate a romance movie script and so I only used scripts from movies in the romance genre. I was hoping that by using scripts from the same genre, my Markov analysis would generate a more coherent movie script. In order to grab all of the romance scripts off of IMSDb, my program first navigates to a page that lists all of the titles in the genre and then copies the corresponding script for each title to a text file that is locally stored in a folder called "scripts." In order to extract the corresponding script from the HTML page, I used the Beautiful Soup package and had the program search for certain markers that indicate the start and end of the movie script. In this case, those markers were "<pre>" and "</pre" (respectively).

This program generated upwards of 200 scripts, many of which had inconsistencies in formatting that would have had consequences on the coherence of my Markov-generated script. My Markov model essentially works by reading each script and creating a dictionary where each key is a word from the script ('prefixes') and each value is a list of the ten most common words that follow it ('suffixes'). In order to generate this list of most common suffixes, I had the program create a nested dictionary that kept track of the number of occurrences of each suffix. I then had it sort that nested dictionary by word frequency and took the top 10 words to insert into the final dictionary used for script generation. I hoped that by limiting the word choice to commonly used words, I would be helping my Markov analysis generate a more coherent script. It is important to note that by doing so, you are sacrificing uniqueness.

In order to generate a script, my program then randomly chooses to start with 'EXT.' or 'INT.' (which are location terms commonly found at the beginning of movie scripts). Suffixes are then randomly chosen based on its prefix word and the program runs until a 500-word script is generated. This word limit is arbitrary.

Results

The majority of my results made absolutely no sense. Despite using movie scripts from the same genre, my model was unable to generate a coherent script. While there were sentences here and there that made sense, the overall script had zero plot. New characters were constantly being introduced and it was rare that a character would have more than one line / monologue in a scene. This lack of consistency meant that there was no buildup in plot or relationship which led to rather random scenes. That all being said, many of my results were very funny. Here are some examples:

> DR. TRAMMEL
> Doctor Rumack, I'm a big deal.
>
> BONNIE
> How did I didn't know how we could just wanted the car in my life to a long way that you get a small room for you get some people will not have been looking to his eyes follow the same room -- a lot of a good thing I was going on.
>
> INT. TOWER ROOM - PRESENT EXT. RUNWAY - THE ROOM TO
> A moment to be with him. They walk away. He doesn't answer. Just then I was in front door. The camera tracks with you. The man and then you think you're gonna get back on the front and they had not going out into an instant of his face of the front seat, her and looks up with you.
>
> JONATHAN
> I'm sure it was so you are still staring into it. You are not in my life in this time he gets his head.
>
> EXT. SCHOOL GIRL.
>
> CHARLOTTE
> What are you are on a man on it. He walks over with an apple red car.
>
> JAMES
> I can't get it and I think I just got to her face.

I believe that the incoherence of my Markov-generated script had a lot to do with the fact that the generation of each word depended solely on the word before it. None of the words before that were taken into account and so rather incoherent sentences were generated. In order to make this model more reasonable, I believe the best next step would be to increase the number of words in each prefix to make phrases. By looking at phrases as opposed to words we might be able to generate a more reasonable script, although we would be sacrificing some uniqueness.

Alignment

I came into this project without having ever scraped the internet or ever having performed any sort of textual analysis. My first priority was to develop my coding / problem-solving skills in these areas. That being said, I was also very interested in understanding text generation (as I encounter it everyday on my phone) and was hoping to gain an understanding of the Markov model from this project.

Throughout this project, I ran into many challenges that opened my eyes up to the complexities of web scraping and text generation and sparked many questions related to improvement. For example, while reading the HTML pages containing movie scripts, I found that there were many inconsistencies in formatting on IMSDb, so it was difficult to create a program that could perfectly capture each script. For this reason, I went through and manually parsed through each script to ensure that certain markers existed, such as "EXT." and "THE END." I ended up having to delete many scripts so my final project only included 50 of the 200 romance movie scripts on IMSDb. While using this database did require some manual work, it was one of the few databases I could find with such a wide variety of formatted movie scripts. Based on the fact that a simple Markov model could actually generate some (though few) coherent sentences from these texts, I believe that this database served my purposes quite well.

While working on the text generation part of this project, I realized the multitude of ways that my Markov model could be made more reasonable and complex. I mentioned earlier that the best next step would be to increase the number of words in each prefix to make phrases but I wonder what other ways text generation models have improved. I also wonder whether generation models can be taught on more than just one type of source. Perhaps I could have trained my model on more than just romance movie scripts and used books or related articles. Perhaps it would have been interesting to generate a movie script using random articles from the internet and a predefined set of script terms / character names.

Reflection

Overall, I am very proud of the work that I accomplished during this project. Although my model was far from perfect, it generated funny results and I learned a lot developing it. I mentioned that I came into this project without having ever scraped the internet or having ever performed any sort of textual analysis. The fact that I did both of those things in the short span of 2 weeks is so exciting! There are definitely many ways that my project could have been improved. For example, I could have increased the number of words in each prefix to make phrases and generate a more reasonable script. While this would have been cool to implement, I did not have enough time at the end of the project to do so. I probably could have scoped my project better but I am happy with what I accomplished during the time that I set aside for this project. I also made a specific point of implementing my own Markov model (as opposed to using Allen Downey's solution) which is also something I am proud of.