Text Mining Project
Markov Analysis of YouTube Descriptions

Jeremy Garcia

Project Overview

For my third mini-project, I used YouTube as my only source. Using the module BeautifulSoup, I was able to gather YouTube descriptions from various videos and combine them into one list. From there, I was able to perform Markov analysis on the combined descriptions and generate a random YouTube description.

Implementation

To implement my project, I searched through the page source code of YouTube videos using the BeautifulSoup module until I found the video description. Then, I combined the video descriptions of multiple videos into one text so that I could perform Markov analysis on multiple video descriptions. Once I had a list with one, combined video description, I was able to generate a dictionary that mapped prefixes to a number of possible suffixes. From there, I generated random text that was based on the Markov analysis that I performed. By increasing the prefix length, the random description made more sense.

Originally, I had planned on performing Markov analysis on YouTube comments instead of the descriptions. However, my attempts at using the YouTube API called gdata were fruitless. Also, BeautifulSoup could not find comments in the YouTube page, since the comments were not in the page source code. Therefore, I decided to modify my analysis for YouTube descriptions of different videos instead of YouTube comments.

Results

For my results, I found that for the Markov- generated descriptions to be good, I needed a lot of different YouTube descriptions that talked about similar things. However, since most YouTube video descriptions are rather short, some of the randomly generated descriptions were either too close to the original descriptions, or too repetitive. To work around this problem, I found the best results were from doing Markov analysis of videos from the same channel or from videos that were about similar things.

The following are examples of 50 word descriptions from videos that mention Donald Trump:

"Trump will win the Nevada Republican caucuses over rivals Sen. Ted Cruz, Sen. Marco Rubio, Gov. John Kasich and Dr. Ben Carson. In conversations with Nevada Republicans ahead of Tuesday's caucus, there seemed near unanimity in support for Donald Trump. CNN's Wolf Blitzer sat down with presidential candidate Donald Trump"

"an interview with Jake Tapper. Analysis of the GOP frontrunner's speech at an Atlanta rally. Donald Trump talks to CNN's Wolf Blitzer about his fellow presidential candidates and what he would do if he's elected. After a victory in South Carolina, Republican frontrunner Donald Trump to discuss a wide range"

Reflections

My main regret for this project was not knowing which projects were possible before I started work on my project. Originally, I planned on working with Facebook and I spent some time trying to get Facebook working until I realized that it was impossible. Also, I spent many hours trying to get the YouTube API gdata to work, but it wouldn't, so in the end I decided to just work with BeautifulSoup which I knew would work.