TextMining

Jared Briskman

February 2016

1 Overview

I used the twitter API to pull the full log of a twitter user's tweets. I then parsed the resultant JSON to get the elements I wanted, cleaning up the text a little. I then used both a prebuilt Markov text synthesis module and one I cobbled together to generate some text from the processed initial dataset.

2 Implementation

I did not implement twitter API calls with OAuth2 in python, because I only needed to make one call. Instead, I just queried their REST api manually, to get an initial dataset. This was a large JSON file, so I parsed through in python to concatenate only the text from each tweet. This plain utf-8 string was then written back to a file. After that, I used an MIT licensed markov text synthesis module, *markov-text-master* to very quickly do something interesting with my data. After that MVP, I tried to put together a markov text generator, and also applied it to the dataset.

3 Results

The seed data used for the text generation came from twitter user "Florida Man", who tweets astounding true eponymous headlines. Due to the nature of these headlines, reader discretion is advised. Output from *markov-text-master* can be seen in "markov-text-master-text.txt, whereas output from"generate.py" can be seen in "OutputText.txt". The latter is rather dissapointing compared to the former, mirroring how I lost a lot of speed when I realized I had reached the limit of this dataset for this tool already.

4 Reflection

There's some clear steps to imporve this project. Writing a much better markov text synthesis engine, or implementing OAuth2 and calls to the twiiter API to generate text based on an arbitrary user are clear directions. Looking back, I was able to get to MVP, being generated markov text, very quickly. However, I didn't get much past that stage. I think a more interesting dataset could have helped push past that stall, like the Hillary Clinton email dump I had on hand, but due to other time constraints, I never got around to implementing it.

I think ultimately I got a fair amount out of this project, but honestly did not have as much time to work on it as I would have liked, even with a day or two extra, due to external factors. All the more reason to budget more time and delve deeper into the next one.