Text Mining Writeup

Nathan Yee

February 2016

1 Project Overview

In this mini project, I made a Tupac lyric generator. This project was done in three stages. First, I used beautifulsoup to scrape lyrics off the web. Second, I used python to parse the text and make a dataset. Third, I utilized Markov chains, the dataset, and python to generate lyrics.

2 Implementation

The structure of the pipeline can be described with four main python scripts.

1. getLinks.py - Given base link. Scrape Tupac song links. Saves link list as links.pickle
2. getLyrics.py - Imports links.pickle. Scrapes lyrics from list of links. Saves lyrics as lyrics.pickle
3. assembleDataset.py - Import lyrics.pickle. Makes the Markov Chain dictionary and first_word list. Saves as prob_dict.pickle and first_word_list_.pickle
4. generateText.py Import prob_dict.pickle and first_word_list_pickle. Uses Markov chains to generate Tupac lyric using the data

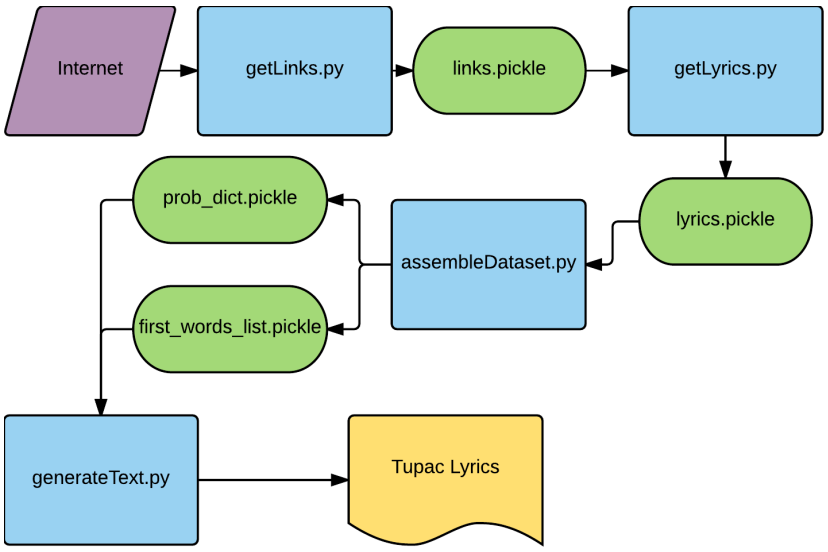

Figure 1: Code Pipeline

3 Results

My accomplishments can be organized into three sections.

1. Web scraping: Used urllib to scrape data from the web. Used beautiful soup functions to parse beautiful soup objects. Ultimately extract lyrics of many Tupac songs.

2. Pickling / Code structure: Utilized pickle to save data to my computer. I then organized my code by creating python scripts that each served a particular purpose. The result is very readable, understandable, and editable code.

3. Markov Chain Algorithms: Used Markov chains to generate the new Tupac lyrics.

Example of randomly generated Tupac song:

I see, that scary, get wreck, ahhhhhhhh-iight? UHHH!
A damn about from enemies deceased
You can handle it feel?
it Cali sticky icky

If I cold inside baby I wanna stay on the way
And to the cash game distorted
There's war on you sleep in ninety-fo'
(Take money) you don't we'll be trouble when I ain't nothin'

No pleasure there's only thing was poorer than tissues
Rather have peace in the years making bucks to the Regal
I love from the S N Double O P Dogg my
Maybe now, I'm a new street fame made her, but a

You done fucked for settin' traps
God sent to you from a getto fantasy, hopin' things will
Gettin high, iced out on me, I recollect we ain't mad
I'm down to ball in my gun so much mail to

4 Reflection

This project taught me the importance of well structured code. Each python script is cleanly written and does a single task. The scripts communicate to each other using stored pickle data. In terms of improvement, I did limited git commits. Increased version history definitely would have helped me. In the future I will commit much more often.