Alexsandr Vobornov (av510)

17 February 2020

Data Science

Project Proposal

I would like to analyze rap and pop lyrics to see if repetition of certain words has any correlation to the average number of plays songs get. To make the script more powerful, I thought about also utilizing a slang dataset that would query for words in addition to the queried phrase (e.g. “money” would also include “bread”, “benjamins”, “bands”, etc.). The slang dataset is available using an Urban Dictionary dataset or by scraping a rap wiki that has a list of terms. To give a clearer example, a user can query a word, and the program would return a table that includes <Song Name, Artist, Word Occurrence, Plays>.

Additionally, I will graph occurrences and the average number of plays per occurrence for that word. For example, this would tell you whether mentioning “money” more would result in a higher or lower number of plays. The machine learning portion of the project could potentially find a trend between occurrences of a word(s) and average plays in the entire dataset, returning a list of verbs and nouns that result in a lot of plays. An alternative application is to determine what an artist’s interests are by comparing mentions of substances, friends/family, money, etc. across their work.

Where will I get the music data?

* <https://genius.com>
  + This website has a top 100 ranking for rap and pop with lyrics, so I could use their API to pull lyric data or attempt to scrape it. All of the data is updated, and you can look up lyrics based on song name.
* Spotify API
  + Useful to get song data (song names, artists, plays) to feed to genius API to retrieve lyrics.
* <https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics>
  + This dataset excludes some artists that I want to analyze, and I don’t think I will have any issues adding more data to it.