Category: Free Topics
Project: Song Lyric Genre Analysis

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.
    a. Just me
        i. Name: Surya Vangala
        ii. NetID: suryav4
    b. I am team captain
2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?
    a. I will make a classifier which will take in the text lyrics of a song and output the genre of the song
        i. I will output a ranking of genre similarity instead of one genre
    b. The task is important/interesting to me because I like music and sometimes songs can be on the border of genres and I want to see what the algorithm says it belongs to
    c. My approach + tools/systems/datasets:
        i. Dataset Creation:
            1. I will go to each genre in spotify, take the top N current artists, take their top M songs, find the lyrics for each song and aggregate into a database of text files for my classifier
                a. N and M are tentatively set to 10 but I will probably need to increase substantially
                    i. I may need to look at different eras of the same genre not just current rankings
                    ii. I may need to consciously select artists in the same genre with different subject matter musical styles
                    iii. I have to compensate for the fact that lyrical content is not a perfect determinant of song genre
                    iv. I will try to preserve differences in accent, pronunciation, ect through deliberate misspellings and grammatical errors in lyric text
                b. Will try to avoid overfitting by selecting as many artists as needed
                    i. Searching for artists lower down on rankings may help with the classifier, I may adjust this
            2. I can augment dataset by searching for datasets of artists on kaggle and elsewhere
            3. Shooting for 80-20 train test split
            4. The model will only receive lyrics no other data
            5. Database will be created manually

ii. Training
    1. My first goal is to train an Okapi BM25 model on the above dataset to be able to classify by genre
    2. I will experiment with n-gram size
    3. I will experiment with running multiple classifiers in parallel and making a ranking system
iii. Test:
    1. I will test the classifier on test data from database above
        a. Songs from other artists
        b. New songs from the same artist
    2. Test data can be aggregated by looking further down rankings or by searching for artists from past years
iv. Frontend:
    1. I will make a local frontend which people who download the repo can run
    2. This will have a text box and will allow users to input their own songs and get a result

d. Expected outcome and work evaluation:
    i. I am shooting for above 50% accuracy on test data. Considering the number of genres out there + timeframe + solo project this seems reasonable
        1. Lyrics are not the only defining characteristic of song genre, so this can be tricky
    ii. I want good results on new songs
        1. Hard to quantify because not part of train or test dataset

3. Which programming language do you plan to use?
    a. Python
    b. Frontend Languages (JavaScript, CSS)
4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.
    a. Data collection
        i. Estimated time: 5 hr
            1. Depends on how much data is needed, may go back to add more data
        ii. I must find train and test data sufficient enough to accurately train a model
        iii. Lots of songs must be found and their lyrics must be aggregated
    b. Data processing and model coding
        i. Estimated time: 10 hrs
            1. I have to decide through trial and error how to make the classifier work. I must chose n-grams, classifiers, and I must develop the code
                a. Other unforeseen problems as well as opportunities to improve code

c. Frontend design
    i. Estimated time: 5-10 hrs
        1. I have never made a frontend for a ML classifier before and I have to learn how to do this and how to integrate with the model