CS 410 Project Proposal
Song Retrieval using Sentiment Analysis

The Group

The group consists of Arnav Jain (Arnavj3) and Nikhil Sahni (Sahni4). The captain of the group will be Arnav Jain.

Our Free Topic

We plan on scraping the web for song lyrics and building an informational retrieval system. We will conduct a sentiment analysis of these lyrics and classify the songs. The user will then be able to query our tool with a combination of parameters like song sentiment, artist, genre, etc. We think this is incredibly interesting as it introduces a novel and useful way of searching for music. An example query for our tool would be "I want a happy country song by ____ artist"

Our planned approach is to first gather the data and clean it into a processable format. We will then work on building a program to conduct sentiment analysis on the gathered lyrics. Once we have a feature set for each song's lyrics, we can start building our search engine that will allow the user to use our tool. We plan on using NLTK for sentiment analysis and text processing tasks. We will scrape and gather data using python. We are still unsure of the data storage tools we will use and plan on making that decision once we start the process of collecting data. This way we will be able to make a more informed decision based on the nature of the data we will work with.

The expected outcome is a working multi-feature search engine for songs. To evaluate our search performance, two useful metrics will be precision and recall. We think a combination of analytical and human oversight will help us evaluate our tool.

Programming Language

We will primarily be using Python for this project. We will be using multiple packages such as NLTK, MeTAPy and Selenium to perform the sentiment analysis, create the search engine and scrape the web

Workload

We expect this project to take us anywhere from 50-60 hours to complete. The division of time is expected to be as follows:
Phase 1: The first phase of this project is to scrape the web to create a dataset of as many songs and lyrics as we can. This would take us 10-15 hours in total.
Phase 2: The second phase consists of creating the search engine to be able to give accurate recommendations of songs only based on the lyrics searched. This would take us 15 - 20 hours
Phase 3: The third phase consists of adding sentiment analysis to the search engine so that users could search for genres and moods and accordingly get song recommendations. This would take us 15-20 hours.