Design Documentation - Assignment 1

Description -

The project aims to help users find the name, information about the song. The user can query using any one or more of the following -

1. Lyrics
2. Artist Name
3. Song Name

Dataset -

A collection of about 1000 songs scraped from genius.com using a custom scraper.

Technologies Used -

Python and it's in-built libraries.
External Python Libraries - requests, beautifulsoup, nltk, numpy, flask

JavaScript.
External JavaScript Libraries - jQuery

Implementation -

1. **Scraping** - The dataset was scraped with a custom scraper using requests, beautiful soup libraries. The songs are saved as JSON files with fields such as artist_name, song_name, lyrics, song_link, etc.

2. **Preprocessing** - The preprocessing step involves generating the terms, bi-words for each song, and storing their frequencies in the same JSON file. We also save the modulus of the weight vector to save computation time on future runs. This is a one-time step which helps reduce search time.

3. **Querying** - The user can enter their query on a web page which is displayed using Flask. The query is then preprocessed in the same way as the documents, and the weight-vector is calculated. This weight vector is used to display 10 documents having highest cosine similarity. The querying is done by sending an async AJAX request to the back-end using jQuery.

Data Structures -

Python Dictionaries, Lists.
Numpy Vectors.