# CS4642 - Data Mining & Information Retrieval
# Song Search Engine - IR Project Report

***Git Repo :*** *https://github.com/UdeshAthukorala/IR_Project-Song_Search_Engine*

This Sinhala Song Search Engine was created by using ElasticSearch and Python.

## Data Description

Data for the search engine was scraped from the sinhalasongbook.com. Data was scraped by using python 'scrapy' framework. Since scraped data contains many fields in the English language, those fields were translated to Sinhala using 'googletrans' python library. This dataset contains 987 Sinhala songs with the data like *title* - song title in both Sinhala and English languages, *song_lyrics* - song lyrics in Sinhala, *views* - view count of the song, *sinhala_artist* - singer's name in Sinhala, *sinhala_lyrics* - Lyricist's name in Sinhala, *sinhala_music* - composer's name in Sinhala, *sinhala_genre* - song type in Sinhala, *english_artist* - singer's name in English, *english_lyrics* - lyricist's name in English, *english_music* - composer's name in English, *english_genre* - song type in English

## Indexing Techniques

In indexing the I used the 'ICU_Tokenizer' which is a standard tokenizer and which has better support for Asian languages to tokenize text into the words. Elastic search 'edge_ngram' filter was used to generate n-grams.

## Querying Techniques

Aggregation technique was used to get aggregated data based on a search query to filter search results based on specific fields. 'Cross fields' and 'Phrase prefix' multi-match queries were used for the querying.

## Advanced Features

Rule-based text mining is used to understand and extract data from the user entered query string. Different lists maintained with the keywords related to artist, writer, and composer. If the query contains a keyword related to only one field, it was removed from the query and did 'phrase-prefix' type query. If it contains keywords related to more than one field did a 'cross-field' type query. If the query contains a number or a keyword related to rating, used the 'range query' or else used the 'faceted query'. Aggregations were used to support faceting.

The search engine supports many types of queries. It supports searching by the title, artist name, writer name, composer name, or using the part of the lyrics. Since I have saved data in both Sinhala and English languages it supports bilingual search. Search Engine can identify ranges given in the search query and sort by view count like '**අමරදේව ගැයු හොඳම සින්දු 10**'. This can identify synonyms related to specific fields. It supports to the wild cards also.