CS4642 - Data Mining & Information Retrieval
IR Project Report
Music Artists Search Engine

Git Repository: https://github.com/HasiniRangana/Music_Artists_Search_Engine

This Music Artists Search Engine was created by using ElasticSearch and Python.

Data Description

The information about music artists were extract from [pinterest.com] (https://www.pinterest.com/musiclk) and Wikipedia. This data set contains more than 100 music artists with the data like artist name, gender, birthday, country, songs, awards, description of that person, and votes which get by artist.

This search engine supports both Sinhala and English language queries. The data were translated to Sinhala using "googletrans" python library.

Indexing Techniques

In indexing I used the 'ICU_Tokenizer' which is a standard tokenizer and which has better support for Asian languages to tokenize text into the words. It can easily be installed in ElasticSearch and for the index.

Querying Techniques

- **cross_fields** - Looks for each word(token) in every field.
- **phrase_prefix** - Looks for whole query in every field.

Cross fields and Phrase prefix are the two types of multi-match queries used. If the query is given specifying one field by using synonyms, then the search will do in *phrase_prefix* type. After removing synonyms, the whole query will be searched in the given field. If the query is given specifying more than one field or not specifying a field, then the search will do in *cross_fields* type. After removing synonyms, each word in the query will be searched in the given fields or every field. Aggregation was integrated with queries to get aggregated data with the search result.

Advanced Features

The search engine supports range queries. As an example "හොඳම ගායකයන් 10" will return the 10 artists having the most number of votes. This can identify synonyms related to specific fields. This search engine supports both Sinhala and English language queries.