

CS4642 - Data Mining & Information Retrieval

IR Project Proposal

Music Artists Search Engine

Scope of the project

In this project, I am going to build a simple Search Engine for Music Artists by using Information Retrieval concepts. And also I hope to use the Sinhala language in this search engine.

Research Objectives

- Creating a Music Artists database by scraping web pages.
- Performing indexing.
- Design a simple application to pass search keywords to search engine, and retrieve the search results to the user.

Introduction

When doing this project, first I hope to create a Music Artists database and I expect to add more than 300 records in there. The dataset will be containing about famous artists with the data like name, country, gender, birthday, songs, genres of music, awards, and description of that person. In querying, search queries with the full text search will be implemented. The system will provide the auto-completion for the queries as well.

Proposed Methodology

The IR engine is expected to be built using ElasticSearch. The proposed methodology consists of four main steps. They are, Web Scraping and preprocessing data, Indexing, Query Modification using the user inputs and retrieving the search results.

1. The first step requires to create a sufficient database by web scraping and web crawling. After scraping it is required to preprocess the scraped data.
2. In the second step, scraped data should index using a search engine.
3. Then it is required to create a simple user interface and an application to take user inputs like search keywords and those user inputs should be preprocessed and should modify to a proper query before entering the search model.
4. Finally, the search engine must run the modified query on the indexed database and retrieve the search results in an ordered manner to the user.