IR Project Proposal

Sinhala Song Search Engine

Introduction

In this project we are going to create a Sinhala song database and a simple Sinhala song search system using information retrieval concepts.

Research Objectives

- Creating a Sinhala song database by scraping web pages.
- Index the scraped song lyrics
- Design a simple application to pass search keywords to search engine, and retrieve the search results to the user.

Proposed Methodology

The proposed methodology consists of 4 main steps. They are, Web Scraping and preprocessing data, Indexing song lyrics, Query Modification using the user inputs and Retrieving the search results.

1. The first step requires to create a sufficient database of Sinhala songs lyrics, by web scraping and web crawling. After scraping it is required to preprocess the scraped data. Song lyrics and song metadata can be stored as a json script.

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.