Index Number:170096L

CS4642 - Data Mining & Information Retrieval
Book Author Search Engine - IR Project Report

Git Repository: https://github.com/dylan96dashintha/SearchEngine_Author.git

This Author Engine was created using ElasticSearch and Python

Data Description

Data for the search engine was taken from ranker.com , peoplepill.com and wikipedia.com
This database contains 102 author records with the data like author_name - names of author in Sinhala and English language, birth_place - birth place of the authors in Sinhala and English language, date_of_birth - date of birth of the author, school - school attended by the authors, book_list - list of books written by the author, about_author - paragraph about the author, language - written languages of the books by the author, category - categories of the books written by the author.

Indexing Techniques

For the indexing part, 'ICU_Tokenizer' is used which is a standard tokenizer and which has better support for Sinhala languages to tokenize text into the words. ElasticSearch 'edge_ngram' filter was used to generate n-grams.

Querying Techniques

Used the sinling tokenizer to tokenize the searched query. Extract the keywords in the search query and identify the related fields and increase the weight of those 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 author name in Sinhala, author name in English, author birth place in Sinhala, author birth place in English ,author birth date and book list.

If the query contains a keyword related to one field such as 'author name in Sinhala', the keyword was removed from the query and did 'phrase-prefix' type query. As the search Engine supports bilingual search, in the previously mentioned example (keyword related to 'author name in Sinhala'), add the 'author name in English' field to the 'final_fields' list and phrase-prefix was done in the fields that include in the 'final_fields' list. If it does not contain keywords, 'cross-field' type query was done.
If the query contains keywords, then the fields related to the keywords are boosted with the caret notation in order to get the most suitable results at the top.

The search engine supports many types of queries. It supports searching by author name in Sinhala and English, by birth place in Sinhala and English, by the birth date of the author, by the school attended by the author, by the name of the books written by the author, by the book category written by the author, or by using sentence which is related to the author.