

Sinhala Song Search Engine

The sinhala song corpus includes over 10,000 songs. Sets of data were collected by scraping the <http://lyricslk.com/> website with the help of sitemap.xml provided by them. Also the <http://sinhalajukebox.org/> is scraped using Selenium ui automation tool. The songs are categorized into 12 categories like oldies, baila, old pop, drama and traditional etc. The attributes of the corpus are,

- 1) Track_id
- 2) track_name_english
- 3) track_name_sinhala
- 4) track_rating
- 5) lyrics
- 6) album_name_english
- 7) album_name_sinhala
- 8) artist_name_sinhala
- 9) duration
- 10) artist_name_english
- 11) artist_rating
- 12) tune_composer
- 13) lyricist

Elasticsearch instance is deployed in AWS ec2 instance and created index called “sinhalasongs”. Indexing has been done for the attributes mentioned above except the numerical fields. Numerical fields have data type of integer in the index mapping. Remaining fields data type is set to text. Simple search query is executed with full text search and limited to 50 results. To organize the results according to most ratings, the sort option is used with normal queries. The landing page is rendered with highest rated songs filtered using range query. Autocomplete search queries as Fig1. with the help of match_phrase_prefix query where slop is set to 3 and max_expansions to 5.

Aggregate (faceting) query is used to display overview of the song corpus. Also the large number of results rendered for the search query are summarized based on the singer using aggregate queries as shown in Fig 2.. Instead of text mining and classification techniques, users were allowed to select tags the search results should be filtered out. Users intents were provided as shown in Fig 3. which leads to higher precision of the results.

Git url <https://github.com/sadilchamishka/Sinhala-Song-Book>

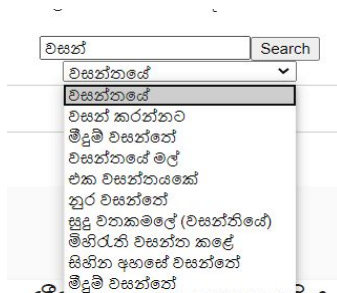


Fig 1. Autocomplete search query

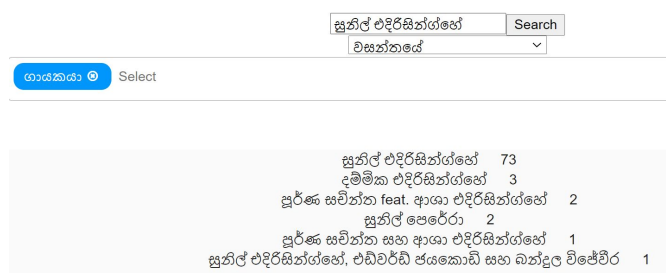


Fig 2. Summarize the search results.

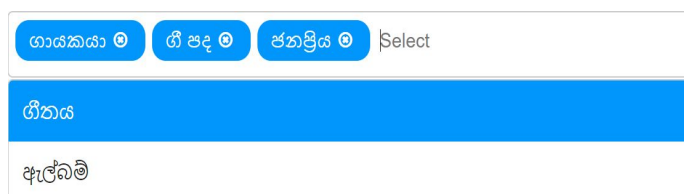


Fig 3. Filter results by multiple tags.