IR System- Sri Lankan Cricketers

The Sri Lankan Cricketers corpus includes 328 records about the Sri Lankan cricketers. The data was collected by scraping the https://www.espncricinfo.com/ website with the help of the Selenium UI automation tool. The attributes(metadata) of the corpus are,

1) සම්පූර්ණ නම 2) උපන් ගම 3)උපන්දිනය 4)වයස 5) පාසල 6) ජායා රූප
7) ජීවන දත්ත 8) අන්වර්ත නාම 9) පිතිකරන විලාසය 10) පන්දු යවන ඉරියව්ව
11) ක්‍රීඩා ඉරියව්ව 12) පුද්ගල වාර්තා 13) කණ්ඩායම් 14) පුද්ගල දක්ශතා

First, an index is created in Elasticsearch called “srilankancricketers”. Indexing has been done for the attributes mentioned above except the numerical fields. Numerical fields have a data types of integers in the index mapping. The data type of the ‘උපන්දිනය’ field is a date type. The remaining fields’ data type is set to text. A simple search query is executed with a full-text search and limited to 50 results. The pagination option is implemented to divide the whole data into smaller parts and the landing page is rendered with the details of the first 9 cricketers by considering the Ids. 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 an overview of the Sri Lankan cricketers corpus. Also, the large number of results rendered for the search query are summarized based on the user requirement (පාසල,උපන් ගම, පිතිකරන විලාසය, පන්දු යවන ඉරියව්ව) 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 the higher precision of the results.
Git url: https://github.com/NadeeshaGarusinghe/Sri-Lankan-Cricketers

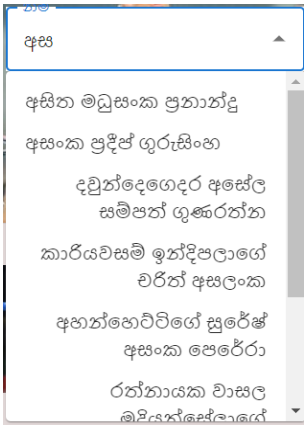

Fig 1. Autocomplete search query

Fig 2. Summarize the search results.

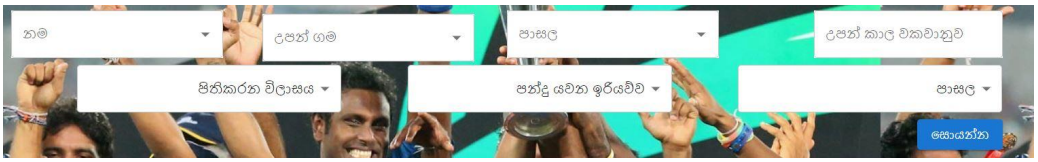

Fig 3. Filter results by multiple tags.