Topic modeling approach for the Automation of Candidate Recruitment process in Sinhala Natural Language

D.D.Weerasinghe

BEng

2020

Abstract

Word shortlisting simply refers to shortening the list. Shortlisting of curriculum vitaes is very time consuming and abstract work. There are number of software tools that can detect and shortlist resumes. Almost all of these tools are available for English language, but similar tools for Sinhala language is not yet available.

There are many attempts of developing language dependent shortlisting tools for languages like Hindi, Chinese, French, Malayalam, Arabic, and English. Most of these tools outperforms the available language independent commercial shortlisting tools as well. Sinhala language being similar to these languages and also being the official language of Sri Lanka along with Tamil, the need of a comprehensive tool to reduce the time for the recruitment procedure is a timely need. Due to the complexity of the language itself the available language independent tools produces very poor results.

This research's main objective is to address the need of a classification tool for Sinhala language to detect and identify similarity of words and rank them according to the weight. A novel algorithm has been developed to detect words correctly and to preprocess then rank then corerctly. The proposed system mainly consists of two stages as text pre-processing and classifying (ranking). Sinhala language resources used in this project were taken from the Language Technology Research Laboratory of University of Colombo and Natural Language Processing Research group of University of Kelaniya. Testing and validations has been carried out by collecting random text samples govenement organizations.

Keywords: *, Sinhala, Natural language Processing, text pre-processing* Publications, Shortlisting, Curriculum vitae