INFORMATICS INSTITUTE OF TECHNOLOGY
in collaboration with
University of Westminster, UK
BEng. (Hons) in Software Engineering

Final year project 2018/2019

Final Thesis

For

SinClassify - Sinhala Text Classification System

By
Anjuka Dulan Koralage
IIT No: 2014022
UOW No: w1534160

Supervised by
Mr. Rathesan Sivagananalingam

Abstract

Since the world beginning, humans, animals use various ways/methods to exchange their ideas between others. From that language is the high-level mechanism to communicate with each other. Language modifies and rich day by day. Language basics laydown on sounds and characters/text. In the recent past, the result of the growth of technology, textual content in web and IT sector was going to a high level. Especially not only vast usage language like English, France, Chinese… but also least usage language like Sinhala, Tamil… the language also shows a reasonable amount of content on Web and IT-related publications. Because of these reasons, automatic text categorization became an important path to many types of research. The propose of this project is to create an accurate text classification mechanism for the Sinhala language. The project called "SinClassifi"

Deeply, SinClassifi project focus machine learning natural language processing path and follow the steps which recommended for textual processing. SinNG5 Sinhala corpus dataset (Lakmali and Haddela, 2018) and more data get from online resources (adaderana.lk and bbcsinhala.com, 2019) combined and remake a proper data set used as the corpus in this project. Because of the limitations, In the data preprocessing stage use customize stop word list by under supervision Sinhala language expertise. TF-IDF use as numerical vectors. For better result, scoped several machine learning classification methods and finally come up with the best one.

The target audience of this project is the Sinhala text users on the web or any other IT-related sector. (Computerized Sinhala text content). Further, they can be University students, Journalists, Sinhala language Researches, normal web readers… The application which builds using the classification, users can copy the Sinhala text content while they refer the document and via a mobile application, they also can classify their set of text.

Keyword – Sinhala language, Automatic text classification, Natural language processing, Multi-class classification