Informatics Institute of Technology

In Collaboration with

University of Westminster, UK

LawE

(Tokenization of legal text for much efficient search results)

A dissertation by

Malik Praveen Perera

Supervised by

Mr. Sudharshan Welihindha

Submitted in partial fulfillment of the requirements for the

BSc (Hons) in Computer Science

Department of Computing

May 2019

Abstract

Law and legal documents have language at its heart, similar to what we find in the languages we use in day to day life but varies with numerous technical and non-technical words used. It is due to these reasons that such a domain with ubiquity and societal importance has not received much attention in the world of Natural Processing Language. There have been certain attempts in trying to use the tools and libraries used for languages in the domain of law and they have succeeded up to a certain degree, but the accuracy levels have not been satisfactory.

The past couple of years there has been some major strides in the improvement of NLP, NLTK libraries and tools etc. specifically targeting the legal domain. LexNLP is such library that tokenizers a verity of keywords, numbers, conditions etc. specifically structured according to the language structure found in legal text.

Yet the world is lacking a proper system where even a user with no prior domain knowledge can easily access structured information from the trillions of unstructured data available in this already overloaded data era.

This research is about structuring a data set so as to be used in a system that can easily be searched and categorized with respective to basic English irrespective of the heavy technical terms found in legal text and documents.

Key Words:

Natural Language Processing

ELK

Text Processing and Indexing

Domain-specific search knowledge

Information retrieval