Document Modeling by Extracting Entities

By

ALICE IRANKUNDA (alice.irankunda@aims.ac.rw)
Supervised by Doctor YABEBAL TADESSE FANTAYE

June 2017

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

Chapter I

1.1. Introduction

In every organization there is a way to communicate ,one of the most popular way to transmit the information is to produce a written report which explains how different activities of the organization are going. For the large organizations there a huge number of reports, imagine the way it is challenging go through each and every report manually. This research has an aim of providing an easy way of visualizing and extracting the important information locked in reports from NGO and large organisations. In 1919, The International Federation of Red Cross and Red Crescent societies (IFRC) has been founded, it has some millions of reports related to humanitarian support,How to know automatically the number of people who suffered from a disease, How to know the fraction of fund spent on shelter ? In this research, There are some solutions to those questions by using combination of statistics and Natural Language Processing (NLP)techniques. Big data and Machine learning is for analysing the huge data by using statistical and computing algorithms. Document modelling by extracting entities is one of the way to deal with natural big data linguistic problems where entity is defined as a single unit of data, it can be classified based on its relationship, Entity can be location , people, organization and so one.

Let R be IFRC report "Afghanistan MDRAF003 26May2016.pdf", it is composed by 12 pages of texts, To extract entities from R is challenging, what are the key points to be performed?

- The sentences which compose a report must be parsed.
- Entities also must be identified in the report
- Relationship between entities must be modelled.

In this research, there is a clear discussion about powerful techniques to answer the previous questions. Natural Language Processing techniques used to sentence level and content based analysis,Natural Language ToolKit (NLTK) for spliting the sentences into tokens and remove the common words and how to work with corpus.The used reports for the implementation of different language algorithms are from IFRC .