Machine Learning Analytics in Social and Health Care using Topic Modelling

(Case Review Application)

Leslie Salami

September 2016

Dissertation submitted in partial fulfilment for the degree of
Master of Science in Big Data (MSc. Big Data)

Computing Science and Mathematics
University of Stirling

Abstract

The purpose of the *Case Review* project is to help create a proof of concept demonstrator for the social and health care industry. This project was fashioned by Nalanda Technology as one of the *Data Lab MSc placement* projects. The Case Review project will look at how to gain greater insight from case documents (*unstructured data*) by using the core *Nalytics Search* and *Discovery* platform. By implementing new analytical capabilities, Nalanda Technology believes that this will help social and health care organizations to deliver more efficient and effective case reviews and interventions.

The objective of this project is to harness the power of *Natural Language Processing (NLP)* in Big Data, by using *Topic Modelling/Extraction* techniques to detect *CIN(Child in Need)* indicators such as illegal drug use, physical abuse, mental illness, neglect and poverty from real case documents of social care workers. This analysis is to help social workers intervene on the child's behalf before it becomes a serious case. These case documents are unstructured and come in different formats such as emails, meeting notes, police statements, and medical reports.

The *CRISP-DM* (Cross Industry Standard Protocol for Data Mining) was used for this Project. The process involved understanding the business problem, understanding the data, preparing the data, creating the model using the *Latent Dirichlet Allocation (LDA)* topic modelling algorithm, evaluating the model (human judgement and computational metrics) and deploying the model (using *Django* as a python web framework) to be used as a proof of concept.

At the end of the project, the model built was able to understand the underlying pattern in case documents and detect CIN indicators. A corpus was built from documents containing case notes, online resources, and 500 serious case reviews. A dictionary was built from the corpus and used alongside the corpus to train the model to create three (3) distinct topics. The combination of these three (3) topics help to detect the required CIN indicators. The final model was saved in Matrix Market (mm) format and loaded by a Django application to be used to analyse real case documents.

The Latent Dirichlet Allocation (LDA) topic model has proved to be a very effective way of analysing case documents such as: detecting CIN indicators, clustering documents and automatically summarizing large volume of unstructured case data. The success of the Project has led to these features being added to the next version of the Nalytics Search and Discovery platform.