MSc Project Report

Cloud Based Tool for URL Metadata Topic Classification

Thulitha Piyasena

2018

A report submitted as part of the requirements for the
degree of MSc Big Data Analytics
at Robert Gordon University, Aberdeen, Scotland

Abstract

With the explosion of web sites and increasing internet traffic is an indication of stronger digital presence of the individual and is becoming a good source of creating personas. To create the persona, it is essential to find the pattern of behavior in the cyber world. In order to identify the virtual behavior patterns, it is essential to perform Topic classification of the sites. Due to the fluid nature of the Internet it is not practical to store all the sites and the topics they belong to and serve rapidly to the demand. Thus, the processing need to be carried out at the point of traffic generation and deriving the Topic classification of the sites before they are browsed.

To handle the requirement, Metadata become the most suited choice to be used for quick classification. Yet due to malpractice in the industry the sites are not built with rich metadata. Stringent preprocessing need be done on the Metadata before they could be used for training to ensure the quality of the model. This project focus on the "Description" Metadata element for training the data set using a cleaning and filtering pipeline which prepare a sufficiently rich data set to be used for model training using LDA model. To maximize the use of these model, a tool need to be available any time, any place on any device. This project proposes a cloud based tool which would scrape the web for Metadata and store it in a NOSQL database for repeated use and classify the sites based on the models trained using the proposed cleansing and filtering pipe line

Key words : Topic Modeling, Metadata, LDA, Cloud, Data cleansing, Data filtering, Dirty Data. Identity privacy,