MSc Project Report

Detection of Hot Topics on Twitter using Named Entities and Event based Incremental Clustering

Mallika Arachchige Prasad Akilendra Jayasumana

2019

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

Abstract

Social media a place where most of the people spend their day to day lives in. It is a place where people communicate and interact with each other, a place where they share their events and get updated on current events and a place where a lot of people are active most of the time.

Many parties will largely value from identifying the current trending hot topics as it will help their business. Marketing companies using trending hashtags when marketing their products are more likely to be noticed. News companies will be able to find the latest news and relevant feedback for them.

While there are many approaches to detect trending topics most of the existing systems have not given much thought to real time performance and have failed to remove unnecessary noise in data making them inefficient.

This research investigates on how to extract events from a twitter stream of data in real time and display them in the form of hot topics. To achieve this an incremental event clustering approach is taken which would be based on the named entities of the tweets. The use of pretrained Doc2Vec generated vectors was proposed to be used for clustering the tweets into their respective events. Additionally, the tweets will undergo a pre-processing stage where noise is removed and an event merging process where similar tweets are merged to the same cluster.

After testing and evaluation phase, the implemented DOH framework gave a Normalised Mutual Information score of 0.911 and a Rand Index of 0.794 after testing it on 100 labelled tweets. The proposed methods and algorithm have proven feasible and given successful results this is justified.

Key words:

Natural Language Processing, Realtime Stream Analytics, Incremental Clustering