Abstract

Extracting data governance information from Slack chat channels

By Simon Quigley

Supervisor: Rob Brennan Assistant Supervisor: Alfredo Maldonado

Masters in Computer Science 2018

Data governance is increasingly important in organisations, and formal systems of data governance have become widespread. Large amounts of communication within these organisations, including data governance information, is carried out over chat channels such as Slack, and this data governance information may not be captured in the organisations data governance system. This dissertation set out to evaluate the effectiveness of a named entity recogniser (NER) in extracting this data governance information from such chat channels. To do this, a dataset of chat messages annotated with data governance related entities was created and a NER optimised and evaluated using this data. Results of this evaluation were promising given the small size of the dataset, but found a high disparity in the named entity recogniser's performance between the annotated entity types. An analysis of factors that affected the NER's performance on each entity type was carried out, finding the type-token ratio and the number of occurrences of an entity type to be a good indicator of the NER's performance on that entity type. The mean and standard deviation of the word length of an entity type were not found to be correlated with the NER's performance. The overall performance of the NER demonstrated that named entity recognition can effectively extract data governance information from chat messages, and future iterations trained on a larger dataset could potentially achieve very high accuracies.