World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:06, 2020

Towards Law Data Labelling Using Topic Modelling

Authors : Daniel Pinheiro Da Silva Junior, Aline Paes, Daniel De Oliveira, Christiano Lacerda Ghuerren, Marcio Duran

Abstract : The Courts of Accounts are institutions responsible for overseeing and point out irregularities of Public Administration expenses. They have a high demand for processes to be analyzed, whose decisions must be grounded on severity laws. Despite the existing large amount of processes, there are several cases reporting similar subjects. Thus, previous decisions on already analyzed processes can be a precedent for current processes that refer to similar topics. Identifying similar topics is an open, yet essential task for identifying similarities between several processes. Since the actual amount of topics is considerably large, it is tedious and error-prone to identify topics using a pure manual approach. This paper presents a tool based on Machine Learning and Natural Language Processing to assists in building a labeled dataset. The tool relies on Topic Modelling with Latent Dirichlet Allocation to find the topics underlying a document followed by Jensen Shannon distance metric to generate a probability of similarity between documents pairs. Furthermore, in a case study with a corpus of decisions of the Rio de Janeiro State Court of Accounts, it was noted that data pre-processing plays an essential role in modeling relevant topics. Also, the combination of topic modeling and a calculated distance metric over document represented among generated topics has been proved useful in helping to construct a labeled base of similar and non-similar document pairs.

Keywords : courts of accounts, data labelling, document similarity, topic modeling
Conference Title : ICAIL 2020 : International Conference on Artificial Intelligence and Law
Conference Location : Toronto, Canada
Conference Dates : June 18-19, 2020