An Extensible Parsing Pipeline for Unstructured Data Processing

Shubham Jain*, Amy de Buitléir**, Enda Fallon*

** Software Research Institute, Athlone Institute of Technology, Athlone, Ireland*

***Network Management Lab, Ericsson, Athlone, Ireland*

sjain@ait.ie, amy.de.buitleir@ericsson.com, efallon@ait.ie

***Abstract*—Network monitoring and diagnostics systems depict the running system's state and generate enormous amounts of unstructured data through log files, print statements, and other reports. It is not feasible to manually analyze all these files due to limited resources and the need to develop custom parsers to convert unstructured data into desirable file formats. Prior research focuses on rule-based and relationship-based parsing methods to parse unstructured data into structured file formats; these methods are labor-intensive and need large annotated datasets. This paper presents an unsupervised text processing pipeline that analyses such text files, removes extraneous information, identifies tabular components, and parses them into a structured file format. The proposed approach is resilient to changes in the data structure, does not require training data, and is domain-independent. We experiment and compare topic modeling and clustering approaches to verify the accuracy of the proposed technique. Our findings indicate that combining similarity and clustering algorithms to identify data components had better accuracy than topic modeling.**

***Keyword*—Unsupervised Data Mining, Information Extraction, Clustering, Topic Modeling**

Shubham Jain is a Software researcher and Lecturer with a demonstrated history of working in the information technology and services industry. . His research interest focuses on Data Science, Machine Learning, Artificial Intelligence, and Software Development applications. Strong engineering professional, pursuing Doctor of Philosophy - Ph.D. focused on Computer Science from Athlone Institute of Technology. Lecturing part-time at Athlone Institute of Technology and UCD Professional Academy.

Dr. Amy de Buitléir joined Ericsson in 2011, where she works as a research scientist. Amy holds a PhD from Athlone Institute of Technology; as part of her PhD research she developed a species of artificial life agents with artificial intelligence that are capable of performing a variety of data mining tasks

Dr. Enda Fallon is the Head of Department in Computer Science at Athlone Institute of technology (AIT), he also worked in as a system architect. In 2003 he founded AIT's Software Research Institute (SRI). Since 2003, Enda has been a principal investigator on over 60 collaborative industry/academic research projects valued at €6.2M. He has published over 70 peer reviewed articles in leading conferences and journals. His research interest focuses on service mediation and adaptation for heterogeneous networking environments.