International Journal of Population Data Science

Journal Website: www.ijpds.org

Clinical Text De-identification and Other Large Scale Processing Tasks in Resource Constrained Environments

Jackson, Richard^{1*}, Dobson, Richard¹, and Stewart, Robert¹

¹Kings College London

Objectives

Clinical text de-identification is a common requirement of the 'enclave' governance model of ethical EHR research. However, there is often little consideration of the engineering task that is required to scale these approaches across the hundreds of millions of clinical documents containing personal identifiers that are resident in the data repositories of a typical NHS Trust. Similarly, natural language processing is an increasingly important field of clinical data science, yet it requires fault tolerant approaches to data processing. This work concerns the development of "turbo-laser" - a distributed document processing architecture based upon the popular 'battle hardened' Spring Batch framework - an industry standard for large scale processing tasks.

MS Office, PDF and scanned documents, and a document re-compiler to deal with EHR legacy issues. Examples of turbo-laser usage include processing 9 million binary documents on modest hardware, within 48 hours.

Conclusion

Turbo-laser is an enterprise grade processing tool, in keeping with the software engineering pattern of 'batch processing' that has been at the forefront of the informatics movement. An open source project, it is hoped that others may contribute and extend its principles, lowering the barrier of large scale data processing throughout the NHS.

Approach

Using Spring Batch, we developed a highly scalable unstructured data processing framework, using the concept of remote partitioning. Remote partitioning allows us to offload processing tasks to any and all computers in a network. With this approach, it is possible to harness the entire compute available of an organisation, whether it be an office of 15 desktop PCs that go unused overnight, or a compute cluster of a thousand processors. This method is especially valuable in the NHS, where the provision of sufficient compute to make large scale analytics possible are often hindered by the lack of available hardware, or difficulties in navigating technical governance policies ill equipped for the demands of modern data science.

Results

Turbo-laser was developed in consideration of the processing challenges common in the NHS. Currently, four types of 'job' are available - De-identification, using the Cognition algorithm, generic GATE output, text extraction from binary files such as

*Corresponding Author:

Email Address: richard.r.jackson@kcl.ac.uk (R. Jackson)