International Journal of Population Data Science

Journal Website: www.ijpds.org

Distributed NLP Framework to create new federated data-sets

Thompson, S[1]

[1]Swansea University

Introduction

Although healthcare systems generate significant amounts of structured data, there remains a untapped wealth of unstructured narrative data. In the UK, 70% of all NHS digital information is in unstructured form. The NHS has no plans to computerise this data, as it is simply would not be cost effective.

Objectives and Approach

Our aim was to make all digitised free text within partner organisations accessible for NLP processing for research, while overcoming information governance challenges. We developed a distributed GATE-based NLP platform enabling NLP models to be automatically distributed and materialised against the free text data in each organisation to create new conventional datasets, which can then be transmitted back using an established governance model.

This work adds NLP capability to the UK's National Research Data Appliances, deployed throughout Wales and beyond and uses many open source components enabling a deployment without additional software licence costs, leading to increased potential use cases.

Results

We have been able to demonstrate a fully federated network of analytical nodes into NHS Wales, which takes the analytical NLP model to the free text data, as opposed to the data having to travel. Under a common, acceptable, governance model, an approval system enables organisations such as health boards to give permission for projects and NLP models to be used against their data.

In a proof of concept project, we have run a number of NLP models over large numbers of documents, which the platform has ingested, converted and analysed.

We have developed a proposal for a common NLP model definition format to enable models to be interchangeable between different research groups and systems. Sharing/discovery of established NLP models is key deliverable.

Conclusion/Implications

The implications of being able to send the query to the data, enables access to this untapped data source, finally enabling the realisation of new datasets, while abiding by any IG framework. The low cost and simplicity will enable a many research opportunities, some of which are already being realised.

http://dx.doi.org/10.23889/ijpds.v3i4.631