Geophysical Research Abstracts
Vol. 21, EGU2019-17507, 2019
EGU General Assembly 2019

Can we derive new value from 200 years of published geoscience?

Richard Jones and Susie Daniels
Geospatial Research Ltd., Dept. of Earth Sciences, Durham, United Kingdom (richard@geospatial-research.com)

An understanding of Earth resources and surficial processes played a fundamental role in the emergence of humans through Quaternary times, and the ensuing dominance of Homo sapiens during the Holocene. Following the age of enlightenment, the development and application of geoscientific methodology underpinned the extraction of iron and coal that fuelled the industrial revolution in the 18th and 19th Centuries. In the 20th Century the exploitation of hydrocarbons was pivotal to the rise of the internal combustion engine, manned flight, international travel, and space technology. Geoscience remains critically important in the 21st Century: even once we have managed to de-carbonise our energy supply, we will still depend heavily on the extractive industries for raw materials in most industrial processes ("if it's not grown, it's mined").

More than two hundred years of systematic research, academic publishing and industrial development have created a vast amount of data across multiple geoscience disciplines. The continual improvement in the price/performance ratio of computer processing, storage and bandwidth, together with a well-publicised increase in the use of data analytics by multinational technology companies, raises the possibility that major new scientific theories might arise from the application of machine learning to our back-catalogue of geoscience data. In some areas of geoscience, such as earth observation data from satellites, this ambition is already being realised. However, in many other disciplines there are significant challenges remaining.

Most available legacy data are in free-text format and are not suitably parameterised for machine learning. We present prototype software that is based on Natural Language Processing (NLP) to analyse and characterise geoscience documents, including published scientific papers in PDF format. Analysis of the semantic content of each text can be carried out in isolation, or in relation to one or more subject-based ontologies. The advantage of the NLP approach is that while the output is rich and multi-dimensional, it is also suitably structured for further analysis (e.g. by other data analytics methods).

A further challenge is that legacy data in the form of geoscience publications often have an important geospatial component that is not easy to access algorithmically. NLP has interesting potential in this regard. Another strategy is the replacement of traditional 'desktop-centric' user interfaces with a 3D geospatial GUI. Recent advances in virtual outcrop technology are an important element of this approach.