Python and ML Algorithms: Automation for Extraction and Compilation of Literature and GIS Data

Taoutaou Aymen*[1], Virginie Morel*[†] , Julien Feneyrol*[‡] , and Claude Cavelius*[§]

[1]Arethuse Geology – Arethuse Geology SARL – France

Résumé

For the past decades, exploration geologists' work has been enhanced and sped up by the easier availability of geoscientific data and digital solutions which have allowed to collect numerous data. The drawback of this trend is that many data are scattered, not organised and in the end not processed nor interpreted. Correctly compiling these data is time consuming for the geologists who need to stay focused on field work, analysis and interpretation. That is why Arethuse Geology, specialised in mineral exploration, in collaboration with DeepLime, expert in python-based geological application, is seeking to optimise data extraction, compilation and export.
Our work aims at developing python-based solutions enhanced by machine learning ("ML") in order to automate the review and analysis of geological and geographical data coming from scientific papers. Specifically, for a given dataset, we target to: (1) scan all data for any duplicate or empty files, rename and sort them out; (2) set-up an efficient interaction with Zotero for literature processing; (3) extract georeferenced data from various sources (*e.g.*, figures, tables, maps, *etc.*). One additional objective is to provide a solution built on top of open-source libraries (such as pandas, scikit-learn, *etc.*).
We have focused our study on Arabian Nubian Shield data from Arethuse collection (_˜1 TB) that encompasses many different file formats and types. Python scripts have been elaborated to automatically and quickly organise a cleansed and sorted GIS architecture in a defined file system structure and create a chatbot linked to Zotero to interact with the available literature within. Most of the GIS data are correctly processed and the outputs of the chatbot show great potential in producing relevant synthetic answers compiling different references. These first results are encouraging, notably saving much time for the geologists, while still opened to improvement through ML. We aim in the future to apply other tools to tackle more complex tasks, such as more complex and advanced data extraction and the integration of Model Context Protocol ("MCP") with the QGIS Python plugin

Mots-Clés: python, machine learning, data exploration & extraction, scientific literature, GIS

*Intervenant
†Auteur correspondant: virginie.morel@arethuse.com
‡Auteur correspondant: julien.feneyrol@arethuse.com
§Auteur correspondant: claude.cavelius@deeplime.io