EGU22-6231
https://doi.org/10.5194/egusphere-egu22-6231
EGU General Assembly 2022

Potential of natural language processing for metadata extraction from environmental scientific publications

Guillaume Blanchy[1], Lukas Albrecht[2], Johannes Koestel[2,3], and Sarah Garré[1]
[1]Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium
(guillaume.blanchy@ilvo.vlaanderen.be)
[2]Agroscope, Zürich, Switzerland
[3]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

Adapting agricultural management practices to changing climate is not straightforward. Effects of agricultural management practices (tillage, cover crops, amendment, ...) on soil variables (hydraulic conductivity, aggregate stability, ...) often vary according to pedo-climatic conditions. Hence, it is important to take these conditions into account in quantitative evidence synthesis. Extracting structured information from scientific publications to build large databases with experimental data from various conditions is an effective way to do this. This database can then serve to explain, and possibly also to predict, the effect of management practices in different pedo-climatic contexts.

However, manually building such a database by going through all publications is tedious. And given the increasing amount of literature, this task is likely to require more and more effort in the future. Natural language processing facilitates this task. In this work, we built a database of near-saturated hydraulic conductivity from tension-disk infiltrometer measurements from scientific publications. We used tailored regular expressions and dictionaries to extract coordinates, soil texture, soil type, rainfall, disk diameter and tensions applied. The overal results have an F1-score ranging from 0.72 to 0.91.

In addition, we extracted relationships between a set of driver keywords (e.g. 'biochar', 'zero tillage', ...) and variables (e.g. 'soil aggregate', 'hydraulic conductivity', ...) from publication abstracts based on the shortest dependency path between them. The relationships were further classified according to positive, negative or absent correlations between the driver and variable. This technique quickly provides an overview of the different driver-variable relationships and their abundance for an entire body of literature. For instance, we were able to recover the positive correlation between biochar and yield, as well as its negative correlation with bulk density.