Graz University of Technology
Institute of Rock Mechanics and Tunnelling

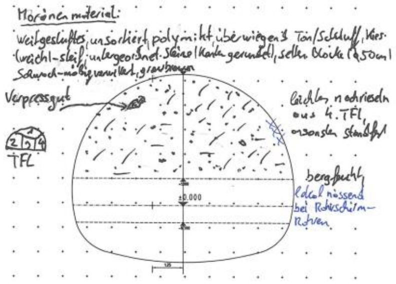
Master Thesis (MT, 30ECTS)

Working Title Incremental learning for extracting text from geotechnical sketches

Description

In geotechnical sketches, the handwritten notes often contain important information that shall be converted to the text (ref. images below) to ensure accurate data analysis based on such images.

The existing methods and out-of-the-box solutions can be used to identify Latin letters and extract them to the text. Unfortunately, the special notations and non-Latin letters are often misinterpreted and therefore the extracted text contains errors and cannot be used further without human check. To by-pass this issue, it is necessary to create a routine for autonomous check of the extracted information's accuracy.


Moränenmaterial:
weitgeklüftet, unsortiert, polymikt, überwiegend Ton/Schluff, Kies (weich-steif), untergeordnet Steine (Kanten gerundet), selten Blöcke (≤50cm) schwach-mäßig verwittert, graubraun
Verpressgut
2 3 4
TFL
Leichtes nachrieseln aus 4. TFL, ansonsten standfest
bergfeucht
lokal nässend bei Rohrschirm-Rohren
±0.000

This work includes the following tasks.

Task 1. Identify a reliable machine learning-based Python library for extracting handwriting to text and re-train it to address letters modified with an umlaut or diaeresis (e.g., ö, ó, etc.).

Task 2. Build a glossary for engineering sketches using available documents and scientific publications.

Task 3. Create a routine for extracting a word from an image in a letter-by-letter base and checking for this word in a glossary, including checking for a misspelling. This solution can be linked to the existing spell-checking routines (e.g., PyEnchant). The routine shall also upgrade a glossary when a new word is detected.

Task 4. Define and implement a metric for the data quality check: this would require a literature review on existing metrics for data quality check and the selection/integration of existing metrics to define the quality of the extracted text (e.g., % of the meaningful words identified correctly).

Task 5. The incremental learning principles (a machine learning paradigm where the re-training process takes place whenever new example(s) emerge) shall be used to create a routine for upgrading a glossary, where accuracy improvement of characters extraction from an image is supported via re-training.

The output of this work will be a routine for building the glossary with demonstrated accuracy improvement for data extraction procedure.

It is desired that a candidate would have at least basic knowledge about machine learning and programming on Python.

Contact Person (s)	Start	Duration	Contact
Alla Sapronova	immediately	6-9 months	alla.sapronova@tugraz.at