DISCOVERING NEAR-MISS OCCURRENCES IN MINERAL EXPLORATION REPORTS

CITS5553 - DATA SCIENCE CAPSTONE PROJECT, CSSE, UWA

July 21, 2020

PROJECT OVERVIEW

1. Project Background and Description

Text data can provide more useful information or knowledge using automated information extraction methods as Natural Language Processing (NLP) techniques advance.

For mineral explorers, past exploration reports are an important resource to understand mineral deposits and their geological and depositional environment in which they form. This may include identifying the rocks that host the mineral deposit, their geological age and the key genetic processes that may be revealed in stratigraphic formations.

Detecting *near miss situations*, where mineral deposit discoveries have not been made but their weak expressions have been recorded, are potentially hidden within mineral exploration reports.

2. Objective

Extracting and visualising near miss occurrences in text reports using NLP techniques. Some trigger words can be: likely contain, potential mineral, or possible mineralisation, etc.

The key challenges are as follows:

- detecting the events;
- classifying them into near miss event, successful event, unsuccessful event, or unknown;
- extracting details/attributes for each event such as location, timing, participants, minerals, ore deposits, tenement, etc.

3. Data

> The WAMEX data contain 30,000+ text reports for geological exploration of mineral resources in Western Australia.

Note: The reports were converted from PDF format to text, so spelling mistakes or joined tokens do exist. There are also sentences in incorrect structure, due to the fact that data were pulled from tables or figures in the original documents.

4. Deliverables

- A report that presents the outcomes of different data exploration techniques applied to your data to meet your project objective, including an effective visualisation technique to help with interpreting the outcome.

- An innovative approach to communicate data science methods to end-users, specifically how to improve the end-user understanding of the analysis process and outcome as well as uncertainty in the outcome – i.e. explainable AI to support end-user decisions.
- Program code (for data exploration and visualisation) on GitHub or BitBucket

5. Affected Parties

Project leader:	Majiga, Email: Majigsuren.Enkhsaikhan@research.uwa.edu.au
End-user:	Paul Duuring, Email: Paul.DUURING@dmirs.wa.gov.au
Students:	TBA

6. Useful Resources

These resources can be useful, though feel free to explore your options or apply your previous knowledge/practice to speed up the implementation.

Key words: event detection, event extraction, entity extraction

- NLP tool: https://spacy.io/

- Graph creation from text:
  Visualisation: https://nlp-tlp.org/text2kg/
  Source code: https://github.com/majiga/text2kg-uwa
  Paper: https://arxiv.org/pdf/1909.01807.pdf

- Good starting tutorials:

https://andrewhalterman.com/post/event-data-in-30-lines-of-python/
https://towardsdatascience.com/natural-language-processing-event-extraction-f20d634661d3

7. High-Level Timeline/Schedule

Week 1: Discussion and data handout

Week 3: Proposal presentation

Week 12: Final presentation

Note: Please send fortnightly progress updates and meeting requests or questions by email to Majiga.

Email: Majigsuren.Enkhsaikhan@research.uwa.edu.au