CITS5553 PROJECT PROPOSAL

DISCOVERING NEAR-MISS OCCURRENCES IN MINERAL EXPLORATION REPORTS

CLIENT: Paul Duuring, Government of Western Australia, Department of Mines, Industry Regulation and Safety

20247459 Ross Michael Green	22665473 Zimin Meng	21140852 Daniel Tang
21726025 Isabella Hardwick	22798485 Ly Khanh Vuong	21515651 Charlie Yin

I. BACKGROUND

The WAMEX data set is a collection of 33824 mineral exploration reports that has been converted from PDF form to manipulatable text in Javascript Object Notation (JSON) format. These exploration reports are an important resource to understand the geography and geology of mineral and ore deposits in Western Australia.

The large repository of text reports and the potential for hidden information and patterns lends itself to automated text extraction methods such as Natural Language Processing (NLP). Some of the most common techniques to extract information from text in NLP include: Named Entity Recognition (NER), Part-of-speech (POS) tagging, event extraction, syntactic parsing, among many others. These techniques can be implemented with a number of different approaches such as pattern matching, machine learning, deep learning, semi-supervised learning, and unsupervised learning.

II. AIM

This project will apply event extraction, classification and visualization techniques to the reports, with the aim of:

- Extracting "near miss" **events** and relevant details from the mineral exploration reports. A near miss event describes an event where mineral deposits have not been discovered but nearby mineralisation is possible.
- Providing an **visualisations** of the information extracted from the reports.
- Organising and **ranking** reports according to the likelihood that they contain information about events.

III. VALUE PROPOSITION

The challenge that the client faces is an overabundance of data in the form of more than 30,000 mineral exploration reports. Each report may contain valuable information about historical exploration and geological data that are relevant for the discovery of mineral deposits in Western Australia. Our team proposes a data science based solution to this challenge by automating the process of information extraction, as well as providing the client with a flexible, understandable, and useful method for visualising the information. The key value propositions we aim to provide include:

- **Cost Savings:** The solution will minimise the degree to which manual report interrogation is required, and instead empower the client with resources, time, and information to drive business decisions.
- **Enhanced Knowledge Extraction:** The solution will allow greater interpretability, reducing the time required to convey important information, and enabling data-driven decision making.
- **Automation, Reproducibility, and Longevity:** The solution will provide the client with the tools necessary to reproduce the results, add future mineral exploration reports, and produce value for the relevant stakeholders after the product has been delivered.
- **Repurposing of Historical Information:** The solution will assist with uncovering potential opportunities without the need for further mineral exploration processes to be performed.

IV. DELIVERABLES

The solution will provide the client with unique and automated means of information extraction from text integrated with a web-based dashboard. This will allow for the enhanced visualisation and analysis of mineral exploration reports and provide the following features for the end-user:

- The ability to select a WAMEX report and perform text-based feature extraction and geological named entity recognition, visualising results alongside the report.
- A table-based display of reports organised by their likelihood for potential mineralisation and/or near-miss explorations, potentially providing functionality to sort by report data such as mineral type, exploration date, location and so on.
- A map-based visualisation tool for an exploration of the information extracted from reports based on location and/or tenement ID.

A data analysis report and presentation that summarises the methods used and the insights gained from the natural language processing pipeline applied to the WAMEX data.

V. METHODS

The work has been divided so that team members working on each section can progress without depending on others, while also allowing each section to iteratively provide feedback and improve. For example, the event extraction algorithm can extract events with only a few trigger words. These events can then be labeled and classified, and the results visualised. As better trigger words and pattern matches are added, the feature extraction process will improve, which will in turn lead to better event classification and report selection.

Preliminary interrogation of the data

Initial interrogation of the data will involve basic text analysis techniques such as the frequency of words, which can be visualised through bar charts and word clouds for exploratory analysis. This will progress to more advanced techniques such as term frequency–inverse document frequency (TF-IDF) to find important keywords. Initial efforts will focus on identifying effective trigger words that will allow the extraction of text that relate to potential "near-miss" events with simple methods. Once the event text is extracted further NLP techniques can be used to classify the events and extract useful information such as the tenements, minerals, etc.

Choice of different data exploration/analytical techniques to test

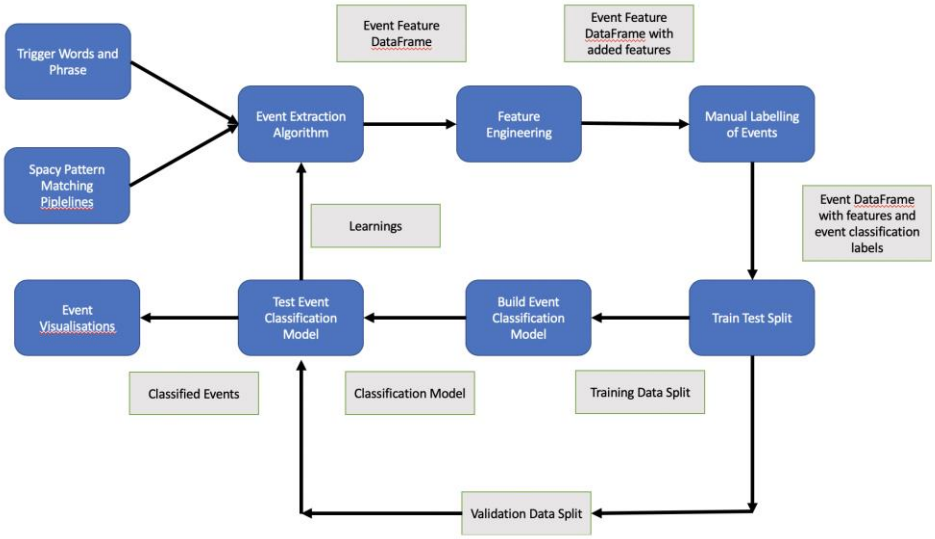
The main areas of event extraction that we have identified will be the use of trigger words to find relevant areas in the text that could relate to an event, as well as the use of pattern matching pipelines, which include more advanced NLP techniques such as named entity recognition, part of speech tagging and dependency linking.

We will then attempt to incorporate the specific geological domain features through transfer learning. This will allow the more general extraction of events or event related features without needing to specifically identify them.

For the classification of events we will explore the use of supervised classification algorithms such as logistic regression, SVMs and decision trees. This will involve the manual labelling of data to train the models and test the results. We will also explore the use of unsupervised learning methods such as topic modelling and clustering to see if events can be naturally clustered into the relevant classifications though their inherent features.

Design and engineer a data prediction pipeline to handle visualisation and analysis of results

A proposal workflow and pipeline for feature extraction and event classification is shown below:



How to make the process understandable and the output usable (e.g. parameter control, incorporating domain knowledge, communicating output uncertainty)

Domain knowledge will be largely captured through the use of relevant trigger words, geological named entities, and grammatical patterns. This will be in a constant state of iterative improvement as we test what features capture events effectively and incorporate feedback from the subject matter expert (SME). The extracted events and the performance of classification models will be validated by the group and the SME collaboratively. This process will inform the team on how to be organise and rank reports according to extracted near-miss events for the client.

Visualisation and organisation of reports to communicate the both processes and results

We will implement a dashboard application with a back-end that is connected to the data set of WAMEX JSON files, allowing us to intuitively visualise information from the text files. Tables will display the metadata and text-based features within a report, and we will aim to work closely with the SME to create visualisations of the information that are easy to use and meaningful to the domain. Interactive map-based visualisations can incorporate information such as the number of events in an area per report, the types of minerals discovered or "nearly missed" by location or tenement, or reports filtered based on their geological location, and so on.

VI. PROJECT DELIVERY TIMELINE (& COSTS)

The team will meet on a weekly basis and has arranged a fortnightly meeting with the client with a total expected commitment of 60 hours per group member (labour costs are specified in hours not dollars). Unexpected costs may occur – i.e. requiring additional data results or to host content on cloud servers, however we will aim for zero cost for the client. A high-level timeline for the key phases of the project is also provided for the client to follow.

VII. PROJECT MANAGEMENT

An additional project management timeline is attached detailing project sub-groups and members, as well as their focus tasks and responsibilities. This management plan has a timeline for all tasks to allow the group to work in parallel – we hope this approach allows our solution to delivered efficiently with time for iterative improvement.

PROJECT DELIVERY TIMELINE

PROJECT TITLE	Extracting Near Miss Events from Mineral Exploration Reports
DATE	18/08/2020

PHASE		DETAILS	UWA WEEK #	1	2	3	4	5	6	7	8	9	B	10	11	12
				JUL		AUG					SEP				OCT		
	PROJECT WEEK:		WEEK START	27	3	10	17	24	31	7	14	21	28	5	12	19
0	Client Meetings	- Initial client engagement and domain-specific research														
		- Recurring progress meetings														
1	Project Conception and Initiation	- Research natural language processing techniques and methods														
		- Project planning and exploration of relevant software packages and tools														
		- Prepare project proposal				due 24/8										
2	Data Preparation	- Extract text chunks using trigger words														
		- Use pattern matching techniques to improve text chunks														
		- Perform named entity recognition on text chunks														
		- Label text chunks to create training set														
3	Model Selection	- Investigate algorithms suitable for NLP tasks														
		- Design and engineer pipeline to train and evaluate models.														
		- Evaluate models on performance and explainability														
		- Apply final model(s) to dataset														
4	Output and Visualisation	- Assess how GeoView data can be integrated with WAMEX text analysis algorithms														
		- Filter and organise reports based on model algorithm outputs														
		- Expand feature extraction and named entity recognition functionality to serve visualisations														
		- Development of interactive and deployable visualisation tool integrated with NLP pipeline														
5	Final Report and Presentation	- Prepare final report													due 23/10	
		- Present results to clients														