**Exploratory Data Analysis**

**Introduction**

This exploratory data analysis report is based on JSON file containing patent documents and their metadata in 2024, acquired from United States Patent and Trademark Office (USPTO). In the further stages of our Final Year Project, we going to use this dataset as a source to fulfil our opportunity statement and project objective of implementing a LLM based Patent Analyzer. The purpose of this exploratory data analysis is to get a feel of the dataset itself, develop understanding of the structure, trends, and relationship that exist using fundamental statistical methods. This step will be significant for us to gain insights on the issues that may arise, formulate ideas on what to do next, and adjust our initial proposed approach if needed. Ultimately, this analysis helps us to be more targeted and efficient in our work of creating a patent analyser as proposed in the project concept.

**Data Collection**

This Exploratory Data Analysis are acquired from United States Patent and Trademark Office (USPTO)’s Application Programming Interface API (<https://developer.uspto.gov/ibd-api/v1/application/publications>) as a JSON file will be processed into a python panda’s data frame of 500 data points. Despite of this, this dataset has a few limitations or biases that may affect our further work such as:

1. Temporal bias
2. Subject matter bias
3. Cultural and Judicial bias

As a result of this dataset being acquired from the USPTO, structure of metadata and the structure of languages use in claims and abstract exhibits judicial jargon and criteria that fits to United States lenses. This may bring us setback when developing analyzer around these texts and reduce accuracy in analysing patents published in different database for example

**Data Preprocessing**

The dataset that we going to analyse has 16 columns and 100 rows. Without further processing, all columns have a string datatype, but as we can see there are columns that needed further processing, either to delete as they are filled with null value or to convert it into the right data type. Additionally, we need to take care of null values we have, 306 values.

**Data Analysis**

As stated in the previous section such as sourcing and preprocessing, this dataset mostly contains string and categorical data types. Some of the string data points has potentials to be a categorical data type (assignedEntityName, AssignedPostalAddressText). Because of this nature, we cannot use basic inference statistic technique such as variance, biases, and t test. Instead, we can use grouping on several columns and do text analysis on the wall of texts.

Here is the analysis divided based on the columns:

1. SubjectmatterCategory

DescriptionText & ClaimText

Both columns are in a long string data type where description usually contains the explanation on the problem statement, opportunity statement, components, structure, features, and usage of the patent product. On the other hand, claims contain boundaries on said product, divided into the main section, independent claim and supporting section, dependent claim.

Because of the nature of the data points, we decided to do a brief text analysis, which we think can give us initial insights and build hypothesis on how to process description and claim which will be essential for summarization and mapping to TRIZ.

The first steps we take are to check whether any null values exist inside these two, if exist, we take care of it. After that, we choose to tokenize texts, get rid of unnecessary stop words and filler words. In this tokenization, we choose to use lemmatization to decrease number of same meaning words, which unfortunately still exist, as we later can see on the word clouds. This part of the analysis gives us 39716 words for claims and 1, 701, 313 words for description across 100 patent documents. Inside these tokenized words, there are 2366 different words overlapping in descriptions and claims. To analyse these tokenized lists further, we decided to create DTM. Several insights we found by doing this are:

1. The matrices are too sparse, for this amount of patent documents (100). To maker it compacts enough, we need to limit number of words to 8, which is not ideal to represent the overall texts of all documents. This imply that the texts are just too diverse, make it tricky to implement conventional word frequency-based mapping method to fulfil our project objective.
2. By extension of this sparsity and word processing method we use, there are a lack of obvious connection between description and claims even for the same kind of subject matter, which is uniform for all our data points.