**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

These biases are occurred because of the data points we extracted are patents designed and filed from 2006-2011, which may not have the same format to the latest publicized patents. This can happen because of change in criteria or changes in trend in how patents document is created, which make patent analyzer build from extraction of outdated data are not accurate.

1. Subject matter bias
2. 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 that is not explicitly unique and do text analysis on the wall of texts.

Here is the analysis divided based on the columns:

SubjectmatterCategory

By using groupby count function, we determined that these extracted patent documents have the same subject matter category which is utility.

FilingDate & PublishDate

These columns give us when a patent is filed, when they are being accepted by US court, and then published into USPTO. We do descriptive analysis on both columns using the describe() function. As you can see the oldest patent is from 2006, with overall mean is from 2011.

Then we decided to convert dates into quarters to zoom out and see overall trend of when patent is filed and when they are published.

Finally, we decided to do calculate the differential days between filing date and published date, from here we can which documents take the most amount of time to be approved and published.

From here we can see that the most days it takes to get approved is 1924 days with 366 days being the average.

assigneePostalAddressText

We analyse this column by extracting the country code of the assignee. We do this to understand the origin of these institutions decided to publish in USPTO.

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. To analyse these tokenized lists further, we decided to create DTM.

Results and findings

From these Exploratory Data Analysis’s findings, we can hypothesize certain nature of dataset that may affect our further work of this project, in our way to create a patent document analyser. Here are insights that we found:

1. Our sample dataset used in the analysis does not necessarily suit what we need to create a patent analyser based on TRIZ. This is because of the fact they are from the same subject matter, are outdated as far as USPTO patent go and limited to only 100 documents. This may give us problem in the training and pipeline development state which may impact the performance of our final product. This is a part of our suspected temporal and subject matter biases stated in the data collection section.
2. Based on the filing date and the publication date column, we found that for this dataset or maybe overall USPTO patent, there is an average of approximately 1 year of waiting time, where the longest may take around 3-4 years. This may indicate that some patents with longer wait time will not be suitable for us to be used as training date as it is considered a lower grade document, but we may need further research with a more diverse data in term of year and subject matter to determine this.
3. Based on the assignee country code, we found that out of 100 patents, most of the company are from United States with China in the second place. This concentration suggests a potential bias towards US-based companies, that might affect skewness of model’s performance inside our pipeline which ultimately will impact our project results presented on the website.
4. There are 2366 different words overlap between claims and description. This indicate there are common semantic used, even though as stated in the next point, this may not be profound as how sparse the DTM is.
5. When we created a DTM, it becomes clear that the tokenized words are too sparse for the size (100 documents), to create a meaningful DTM, we need to limit words to 8 which is not accurate representation of the texts. The large variation create difficulty in recognise pattern and overlap based on the frequency alone. This is not a good sign, even though the number of documents is still limited, as stated on the data collection, all of them are from the same subject matter category.
6. By extension of the overlap and sparsity, we found there are a lack of connection between claims and descriptions even though they are sourced from the same group of documents.
7. There are certain grammars that are widely used in these documents’ description and claims. Documents have grammar that list out how a patent works and claims has grammar with authoritative nature which implied boundaries of the patents.