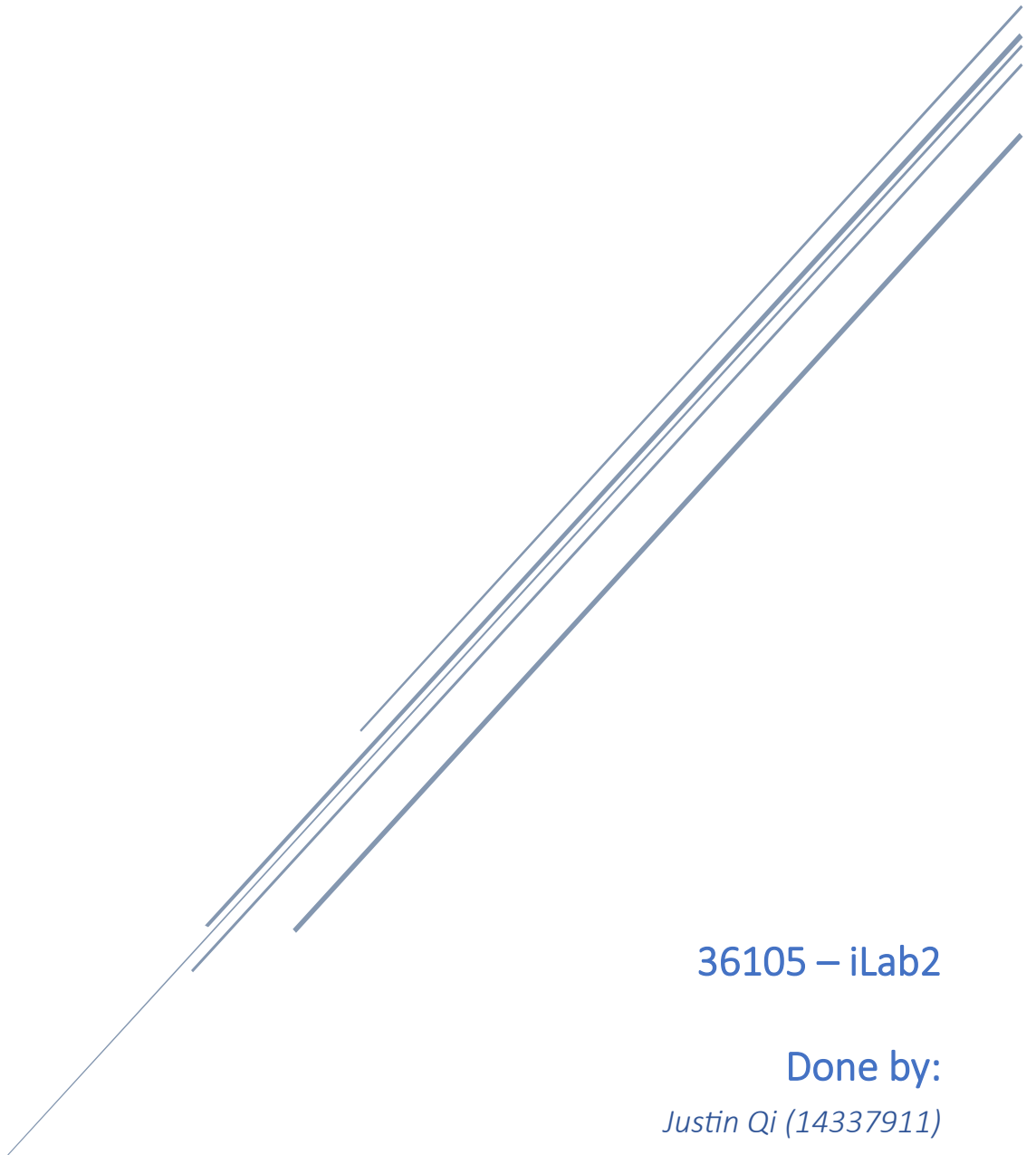


NEW SOUTH WALES DEPARTMENT OF PREMIER AND CABINET

Strategic Intelligence Gathering Engine



36105 – iLab2

Done by:

Justin Qi (14337911)

Phoebe Jacinda-Santoso (24600088)

Antonio Wang (12972042)

Contents

| | |
|--|----|
| 1. Executive Summary..... | 3 |
| 2. Introduction | 4 |
| 2.1 Background and Context..... | 4 |
| 2.2 Problem Statement | 4 |
| 2.3 Project Vision | 4 |
| 2.4 Stakeholder Involvement | 4 |
| 3. Project Objectives & Scope | 5 |
| 3.1 Primary Goals..... | 5 |
| 3.2 Directional Pathways..... | 5 |
| 3.3 Deliverables..... | 5 |
| 4. Client Description and Background Context | 6 |
| 4.1 Client Overview | 6 |
| 4.2 Need for Strategic Intelligence | 6 |
| 4.3 Partnership's Genesis..... | 6 |
| 5. Data Landscape | 7 |
| 5.1 Overview | 7 |
| 5.2 Data Source & Acquisition..... | 7 |
| 5.3 Data Breakdown & Attributes | 7 |
| 5.4 Data Integrity and Quality | 8 |
| 6. Data Pipeline | 9 |
| 6.1 Data Extraction..... | 9 |
| Instructions | 9 |
| 6.2 Azure Blob Storage – Initial Staging | 10 |
| Instructions | 10 |
| 6.3 Azure Databricks Processing: | 10 |
| a. Bronze Layer:..... | 10 |
| b. Silver Layer: | 10 |
| c. Gold Layer: | 10 |
| Instructions | 10 |
| 6.4 Visualization with Power BI:..... | 11 |
| 7. Methodology..... | 13 |
| 7.1 Overview | 13 |
| 7.2 Analytical Framework – CRISP-DM..... | 13 |
| 7.3 Data Science Methods and Techniques | 13 |

| | | |
|------|---|----|
| a. | Longformer to Roberta Model: | 14 |
| b. | Sentiment Analysis and Text Categorization: | 14 |
| c. | Text Classification for Different Aspects: | 14 |
| d. | Text Summarization using T5: | 15 |
| e. | Latent Dirichlet Allocation (LDA) for Topic Modeling: | 15 |
| f. | Time Series Analysis for Temporal Trends | 15 |
| 7.4 | Rationale | 16 |
| 8. | Findings and Values | 17 |
| 8.1 | Insights into Technological Trajectories | 17 |
| 8.2 | International Competitiveness | 17 |
| 8.3 | Visualization through Power BI Dashboard | 17 |
| 8.4 | Value Proposition | 18 |
| 9. | Challenges & Solutions | 19 |
| 9.1 | Constructing the Data Pipeline | 19 |
| 9.2 | Navigating Diverse Data Sources | 19 |
| 9.3 | Facilitating Non-Technical Accessibility | 19 |
| 10. | Recommendations & Future Strategy | 20 |
| 10.1 | Enhancing the Intelligence Engine | 20 |
| 10.2 | Expanding the Power BI Dashboard | 20 |
| 10.3 | Leverage Enhanced Azure Capabilities | 20 |
| 10.4 | Continuous Feedback Loop | 20 |
| 11. | Conclusion | 21 |
| 12. | Appendices | 22 |

1. Executive Summary

Project Overview: Recognizing the technological challenges and opportunities in Australia, especially within Sydney, our team at the University of Technology Sydney collaborated with the NSW-DPC to develop the Strategic Intelligence Gathering Engine Project.

Objective: The primary goal was to design a robust intelligence engine that integrates data from multiple sources, providing actionable insights that can guide NSW's future strategies in technology and market positioning.

Approach:

- **Data Collection:** Initial focus was placed on data acquisition from the LENS API and other relevant sources.
- **Analytical Framework:** Leveraging machine learning models, the data underwent several layers of analysis. The process included text data processing, pattern recognition, and forecasting, resulting in a comprehensive understanding of the existing trends and potential future trajectories.
- **Pipeline Design:** The project boasts a streamlined and automated data pipeline that ensures efficient processing and timely updates.
- **User Interface:** Insights derived from our analysis are presented through an interactive PowerBI dashboard. This user-friendly interface aids in intuitive decision-making by offering a clear view of the data and its implications.

Value Proposition: The project offers a clear framework for strategic planning. By consolidating vast data streams and presenting them in an actionable format, the intelligence engine aims to position NSW at the forefront of technological advancements. The system emphasizes efficient data processing, actionable insights, and scalability.

Challenges Overcome: The project's development encountered challenges, especially when integrating tools like Azure's ecosystem, Databricks, and Power BI. Resource limitations presented obstacles, but through collaborative problem-solving and efficient resource allocation, we created a stable and scalable solution.

Future Prospects: With the foundational systems in place, there's an opportunity for further enhancement. The resources and expertise of NSW-DPC can be harnessed to refine models, incorporate larger datasets, and ensure the engine remains responsive to evolving technological landscapes.

Detailed Analysis: The following sections offer an in-depth look at our methodologies, findings, and recommendations, providing a comprehensive understanding of the project's scope and potential impact.

2. Introduction

2.1 Background and Context

The New South Wales Department of Premier and Cabinet (NSW-DPC) is at the forefront of driving strategic initiatives for the state. As the technological landscape evolves, there's a growing need for a robust strategic intelligence mechanism to navigate the complexities of global patent trends. This report delves into our partnership's genesis with the NSW-DPC, emphasizing the critical need for strategic intelligence in today's dynamic environment. By understanding the client's core objectives and the challenges they face, we aim to provide a comprehensive solution tailored to their specific needs.

2.2 Problem Statement

In the vast sea of global patent data lies the potential for unparalleled insights, but harnessing this potential is no small feat. The pressing questions for NSW, especially its flagship city Sydney, revolve around its standing on the global stage: Are we pioneers, followers, or potential disruptors in emerging technological realms? Tackling these questions demands an innovative tool that can seamlessly translate voluminous data into actionable, clear-cut insights, facilitating informed policy decisions and visionary strategies.

2.3 Project Vision

Our vision for the *Strategic Intelligence Gathering Engine Project* was clear: Develop a tool that bridges the gap between raw data and strategic decision-making. Our ambition was to provide the NSW-DPC a narrative beyond mere data points or visuals. We aimed to illustrate NSW's current global position, its potential trajectory, and the strategic pathways to etch its name in the annals of global tech leadership.

2.4 Stakeholder Involvement

Collaboration was at the heart of this venture. With guidance from the esteemed Passiona Cottee and Scott Perugini Kelly from NSW-DPC and mentorship from Christopher Mahoney at UTS, our project was a confluence of expertise, experience, and enthusiasm. Their invaluable feedback and insights were pivotal in shaping the project's direction, ensuring it remained aligned with the overarching objectives of the NSW-DPC.

3. Project Objectives & Scope

3.1 Primary Goals

Success hinges on defining objectives that align with a project's foundational purpose. In our initial consultations with the NSW-DPC, the blueprint of our project took form with an emphasis on harnessing the power of available data to its fullest potential:

1. **Strategic Intelligence Engine:** The centerpiece of our endeavors was the design and development of a state-of-the-art data processing mechanism. This engine was envisioned to seamlessly integrate data from disparate sources and transform it into actionable insights.
2. **Insightful Analysis:** The goal here was not just data interpretation but a profound analysis aimed at discerning trends, capturing subtle nuances, and detecting overarching patterns. The end objective was to offer a holistic view of Australia's and, more specifically, NSW's stature in the global tech panorama.
3. **Interactive User Interface:** Our commitment was to relay insights in an engaging, comprehensible manner. The adoption of PowerBI was a deliberate move, envisioned to provide decision-makers with a potent tool to derive insights effortlessly.

3.2 Directional Pathways

Our journey was mapped in three meticulously planned phases, each aligning with the overarching goals set by the NSW-DPC:

1. **Data Ingestion:** The cornerstone of our project was to tap into relevant data sources. The LENS API emerged as a primary reservoir, but our approach was holistic, ensuring we cast a wide net to capture comprehensive data.
2. **Transformation & Enrichment:** Raw data, while valuable, needs refinement. This phase was a crucible where data underwent cleansing, processing, and enrichment. The objective was to mold it into a form ripe for in-depth analysis, reflecting the needs and ambitions of the NSW-DPC.
3. **Visualization:** With enriched data at our disposal, the next challenge was its presentation. The Power BI dashboard was crafted to be an interactive interface, ensuring that stakeholders could navigate, interpret, and action insights with utmost ease.

3.3 Deliverables

Our endeavor aimed to equip the NSW-DPC with:

- A robust intelligence engine, architected for routine data updates and scalability to accommodate emerging data streams.
- A user-centric Power BI dashboard, fine-tuned to resonate with the strategic imperatives of the NSW-DPC.
- Comprehensive documentation that demystified the project's nuances, fostering transparency and serving as a roadmap for future enhancements or iterations.

4. Client Description and Background Context

4.1 Client Overview

The NSW Department of Premier and Cabinet (NSW-DPC) stands as a beacon of central governance, with an unwavering commitment to lead, coordinate, and invigorate whole-of-government policy-making and execution. Their mission to address contemporary challenges and forecast future ones has always been rooted in data-driven decision-making.

The Premier's Department's endeavour to remain in continuous dialogue with the community, leading the NSW public sector towards a more inclusive and impactful governance, demands a deep dive into data that reflects the technological pulse of the nation and the world.

4.2 Need for Strategic Intelligence

In an era where information is abundant but insights are rare, the NSW-DPC seeks to harness the power of data to chart out NSW's trajectory in emerging technologies and to ascertain its global competitiveness. To anchor policies and strategies that propel NSW to the forefront of technological advancements, the Department requires a clear understanding of how it stands relative to global benchmarks.

The envisioned Power BI dashboard is not just a tool; it represents a compass pointing towards uncharted territories, guiding the state's investments, policy adaptations, and areas of focus.

4.3 Partnership's Genesis

Aligned with their vision, the NSW-DPC's collaboration with our team aimed at sculpting an engine that bridges the gap between raw patent data and strategic insights. The final deliverable, an interactive dashboard, is anticipated to serve as a pivotal decision-support tool, enabling the Department to structure the future roadmap for NSW's technological landscape and inject enhancements into existing industries.

5. Data Landscape

5.1 Overview

Patents serve as a direct reflection of a nation's innovation landscape, and thus, provide a rich source of insights into technological advancements, R&D focuses, and potential areas of growth. It was with this understanding that our team chose to delve into patent data as our primary data source.

5.2 Data Source & Acquisition

We anchored our data extraction around the LENS API, which, for the uninitiated, offers a robust and comprehensive view into the world of patents. Spanning a period of five years (2019-2023), our data encapsulates:

- **2019:** 2,772 entries
- **2020:** 2,874 entries
- **2021:** 2,725 entries
- **2022:** 2,609 entries
- **2023:** 1,964 entries

5.3 Data Breakdown & Attributes

To ensure a holistic view, we curated a list of fields from the extracted patent data, pivotal for our analyses:

1. Lens ID: The unique identifier for each patent.
2. Jurisdiction: The patent's geographical jurisdiction.
3. Publication Reference: Key publication details.
4. Application Reference: Patent application specifics.
5. Invention Title: The title given to the patented innovation.
6. Inventors: Names of the innovators behind the patent.
7. Applicants: Names of individuals/entities applying for the patent.
8. CPC Classifications: A taxonomy for classifying the subject matter of patents.
9. Cited Patents: References to other patents.
10. Abstract: A brief summary of the patent.
11. Claims: A detailed list of claims made by the patent.
12. Patent Status: Current legal status of the patent.

13. Owners: Individuals/entities that own the patent rights.
14. References Cited: Relevant literature or prior patents referred during the patent's formulation.

5.4 Data Integrity and Quality

Given the importance of ethical data practices, especially in projects that could influence policy and business decisions, we've been meticulous in our approach:

- **Data Source Integrity:** The LENS API is a reputable source of patent data, ensuring the data's authenticity and reliability.
- **Transparency:** All processes, from data ingestion to analysis, are documented to provide NSW-DPC with a clear view of our methodologies.
- **Data Security:** Utilizing Microsoft Azure services, known for their robust security protocols, underscores our commitment to data protection.
- **Open-Source Adherence:** By leveraging the LENS API, an open-source platform, we ensure compliance with licensing and data usage terms.
- **No Bias:** Our analyses and models are designed to be objective, free from any biases that might skew the insights.

6. Data Pipeline

Our data pipeline has been meticulously designed to ensure the highest level of data quality, transformation efficiency, and integration, leveraging state-of-the-art tools and methodologies.

This structured, multi-layered approach to our data pipeline ensures not only the efficiency and accuracy of our data transformations but also the reliability and repeatability of the entire process. It provides stakeholders with the confidence that the insights gleaned from our analysis are grounded in solid, methodical processing and genuine data-driven discoveries.

Below is a step-by-step breakdown of the entire workflow:

6.1 Data Extraction

Using a tailored Python script, we pull raw data directly from the Lens API. This data, primarily in JSON format, contains intricate nested structures, which are valuable but require careful processing to harness effectively.

Instructions

Change API to your desired source

```
API_URL = 'https://api.lens.org/patent/search'  
BATCH_SIZE = 100  
DELAY = 1
```


Define what variables you want to extract from the API, list of variables can be found on Lens.org.

```
# Fields to be included  
include = [  
    "lens_id",  
    "jurisdiction",  
    "doc_number",  
    "kind",
```

Set the year of the data you want to extract from (this is because Lens only allows 10k patents to be extracted in one execution), anything more than that Lens will terminate the script automatically.

```
all_results = []  
offset = 0  
YEAR = 2023 # You can change this each time  
OUTPUT_FILE = f"raw_output_{YEAR}.json"
```

You should then see a raw_output.json file in your directory.

 raw_output

6.2 Azure Blob Storage – Initial Staging

Upon extraction, the raw data is immediately uploaded to Azure Blob Storage. This acts as our primary staging area, ensuring we have a robust backup of the original data and can readily access it for further processing.

Instructions

Create Azure account – Create resource group – Create Storage Account – Create your containers (I named mine 'bronze', 'silver', and 'gold')

6.3 Azure Databricks Processing:

a. Bronze Layer:

The first stage within Azure Databricks involves taking the raw JSON data and processing it. Given the nested nature of the data, we employ sophisticated flattening techniques to "unnest" its contents. The end product of this layer is a cleaned and structured Spark dataframe, primed for further transformations.

b. Silver Layer:

The structured data from the bronze layer undergoes additional transformations in the silver layer. Here, individual tables are merged to produce a consolidated dataframe, paving the way for more advanced and integrative analytics.

c. Gold Layer:

At this pinnacle layer, we leverage external data, notably the CPC classifications sourced from the official CPC website. By mapping these classifications onto our merged dataframe, we enrich the dataset with valuable context and categorizations. Additionally, the gold layer is where our primary analysis is conducted, resulting in the creation of our final dataset, chock-full of insights and discoveries.

Post-analysis, the final dataset is saved back into the gold container in Azure, where it's primed for visualization and further interrogation.

Instructions

We've created an automated workflow on Databricks, you can use the same format as ours. You can configure the frequency of this workflow process and other details. This method is simple and automated.



6.4 Visualization with Power BI:

With the data processed and stored in Azure's gold container, we use Power BI to directly retrieve it. Power BI's robust capabilities allow us to craft detailed, interactive dashboards that present the data's insights in a clear, concise, and actionable manner.

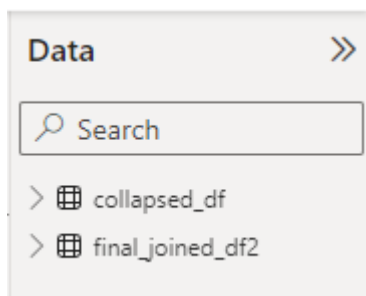
To view Power BI dashboard interface, follow the instructions below or look for screenshots in the appendix section of the report.

Due to the restrictive nature of our Power BI student account, we are unable to share the Power BI service (browser version) to people outside of UTS. To resolve this issue, we might need to consult the IT department of UTS.

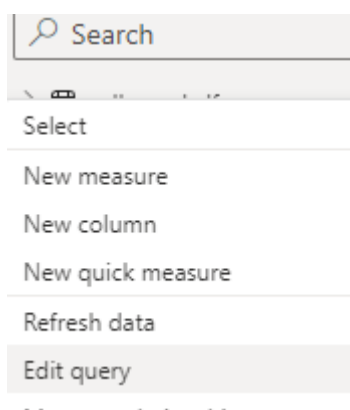
Instructions:

When you open the .pbix file. You will be redirected to Power BI desktop, and you will be prompted that the data source cannot be found. This is because the data source is currently only mapped to our local address. You can change it to Azure, but unfortunately we had difficulty reading directly from Azure and could not troubleshoot this issue.

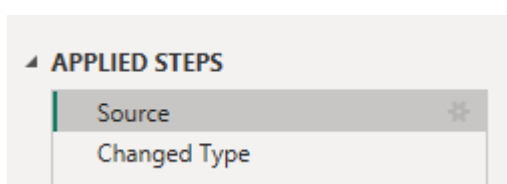
First, at the right side of the screen, you will see a 'Data' side window.



Right click on the first data source, select 'edit query'



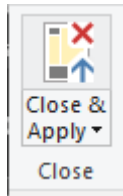
At the right side of the query window, select 'Source' under 'Applied Steps'



Change the source to where the data is located in your computer.

```
Contents("C:\Users\phoeb\Downloads\collapsed_df.parquet"),
```

Click 'Close & Apply' on the top left hand corner.



Repeat the same process for the second data set and the entire dashboard should be working in a few seconds of buffering time.

| | | | | | | | |
|------|----------------|---------------|------------|------------|-------------------|--------------|---------|
| Home | Topic Scanning | Date Scanning | Domain Viz | Region Viz | Aus vs. The World | Key Insights | Summary |
|------|----------------|---------------|------------|------------|-------------------|--------------|---------|

NSW TREND INSIGHTS

| Scanning | Viz | Summary |
|--|---|--|
| Scanning the journals based on topic, date, and origin | Visualisations of the journals' information | Gaining Insights from the curated journal contents |



7. Methodology

7.1 Overview

The nature of our project necessitated a blend of traditional data analysis methods and advanced machine learning techniques. The goal was to translate the extensive patent data into actionable insights.

7.2 Analytical Framework – CRISP-DM

CRISP-DM Approach: Recognizing the importance of a structured approach to data science, we adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which is renowned for its effectiveness in guiding projects from conception to deployment. The stages include:

- **Business Understanding:** By liaising closely with NSW-DPC, we gained a comprehensive understanding of the project's objectives and the specific insights desired.
- **Data Understanding:** This phase allowed us to familiarize ourselves with the intricacies of the patent data, identifying potential areas of exploration based on the provided fields.
- **Data Preparation:** Processing and organizing the raw data were paramount. Given the data's complex nature, we invested time in flattening the JSON files into data frames that would be more amenable to analysis.
- **Modelling:** With the foundational data groundwork laid, we will embark on modelling, aiming to utilize machine learning techniques to extract deeper insights from the data.
- **Evaluation:** Any insights and models developed will be rigorously evaluated for accuracy and relevance.
- **Deployment:** The end-goal remains the Power BI dashboard, which will serve as an interactive platform for the NSW-DPC to access and interpret the derived insights.

Adopting CRISP-DM ensures a standardized, effective approach, enhancing the project's credibility and the reliability of its outcomes. Coupled with our strong ethical stance, we aim to provide the NSW-DPC with insights they can trust, laying the groundwork for informed, future-centric decisions.

7.3 Data Science Methods and Techniques

In this project, we employed a series of advanced data science techniques to extract, process, and analyze our dataset. Our methodology is rooted in leveraging state-of-the-art NLP models and tools from the **transformers** library, ensuring the extraction of meaningful insights.

These methodologies were chosen based on their suitability to handle the intricacies of the dataset and to derive actionable insights. By integrating these advanced techniques, we ensured robust data processing, resulting in an enriched analytical outcome.

Below is a detailed breakdown of the methods and models used:

a. Longformer to Roberta Model:

- **Purpose:** Transformation of text data.
- **Method:** Leveraged the **LongformerModel** and **LongformerTokenizer** from the **transformers** library.

```
from transformers import LongformerModel, LongformerTokenizer
model_name = "patrickvonplaten/longformer2roberta-cnn_dailymail-fp16"
tokenizer = LongformerTokenizer.from_pretrained(model_name)
model = LongformerModel.from_pretrained(model_name)
```

b. Sentiment Analysis and Text Categorization:

- **Purpose:** To extract sentiment from textual data and categorize sentiments into various predefined categories.
- **Method:** Initialized a sentiment-analysis pipeline and subsequently split the text into chunks to better handle sentiment categorization.

```
from transformers import pipeline
nlp = pipeline("sentiment-analysis")
```

The sentiments were further mapped to categories such as "Trust", "Opaque", "Empowerment", etc., based on the sentiment value (POSITIVE/NEGATIVE).

c. Text Classification for Different Aspects:

- **Purpose:** To classify text based on multiple attributes like **Time Horizon**, **Opportunity-Challenge**, **Likelihood**, etc.
- **Method:** Initialized another sentiment analysis pipeline and subsequently classified the text based on the sentiment detected. Categories such as **Time Horizon**, **Opportunity-Challenge**, and others were assigned based on whether the sentiment was POSITIVE, NEGATIVE, or NEUTRAL.

```
from transformers import pipeline
# Initialize sentiment analysis pipeline
sentiment_pipeline = pipeline('sentiment-analysis')

def classify_text(text):
    result = sentiment_pipeline(text)[0]
    label = result['label']
    score = result['score']

    # Mapping sentiment to categories
    if label == 'POSITIVE':
        return {
            'Time Horizon': 'Medium term',
            'Opportunity-Challenge': 'Some opportunity',
            'Likelihood': 'Almost Certain',
            'Degree of Change': 'Medium Change',
            'Scale of Impact': 'Society'
        }
    elif label == 'NEGATIVE':
        return {
            'Time Horizon': 'Short term',
            'Opportunity-Challenge': 'No opportunity',
            'Likelihood': 'Uncertain',
            'Degree of Change': 'Small Change',
            'Scale of Impact': 'Individual'
        }
    else:
        return {
            'Time Horizon': 'Long term',
            'Opportunity-Challenge': 'Unclear',
            'Likelihood': 'Uncertain',
            'Degree of Change': 'Unclear',
            'Scale of Impact': 'Unclear'
        }
```

d. Text Summarization using T5:

- **Purpose:** To generate concise summaries of lengthy textual data.
- **Method:** Leveraged the **T5** (Text-to-Text Transfer Transformer) model for the summarization task. Given the length constraints of the T5 model, we tokenized the input text and then used the model to generate summary outputs. The first 30 rows of the data frame were summarized as a demonstration.

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
tokenizer = AutoTokenizer.from_pretrained("t5-small")
model = AutoModelForSeq2SeqLM.from_pretrained("t5-small")
```

e. Latent Dirichlet Allocation (LDA) for Topic Modeling:

- **Purpose:** To identify the underlying topics present in the textual data.
- **Method:** The Latent Dirichlet Allocation (LDA) model was employed for topic modeling, which is a generative probabilistic model that allows sets of observations to be explained by unobserved groups. The idea is to explain why some parts of the data are similar by positing that they arise from the same underlying topic.

Once the LDA model was trained on the vectorized data, it was used to transform the data to get the topic distribution for each document. This distribution provides the probability or likelihood of each topic being present in a particular document.

```
# Get the topic distribution for documents
doc_topic_dist = lda_model.transform(data_vectorized)
```

After obtaining the topic distribution, the **dominant_topic** for each document was determined by identifying the topic with the highest probability for that document.

```
# Assign dominant topic for each document
filtered_df['dominant_topic'] = doc_topic_dist.argmax(axis=1)
```

By doing this, each document was effectively labeled with the most relevant topic, providing a concise representation of its primary content.

f. Time Series Analysis for Temporal Trends

- **Purpose:** To capture the evolving trends in patent publications over time and forecast future trajectories.
- **Method:** The ARIMA (AutoRegressive Integrated Moving Average) model was employed for time series forecasting, which is a renowned statistical method for analyzing and forecasting time-ordered data. The model captures:
 - **AutoRegressive (AR) patterns:** This component identifies the relationship between an observation and a number of lagged observations (previous time points).
 - **Integrated (I) patterns:** This captures the cumulative sum of differences to make the time series stationary.

- **Moving Average (MA) patterns:** This identifies the relationship between an observation and a residual error from a moving average model applied to lagged observations.

Once the ARIMA model was trained on the dataset, it was used to make predictions, yielding insights into potential future trends in patent publications. This forecasting can offer a forward-looking perspective, allowing stakeholders to anticipate and prepare for upcoming shifts in the innovation landscape.

```
# Define the ARIMA model
model = ARIMA(patent_count, order=(5, 1, 0))
model_fit = model.fit()

# Predict future patent counts
predictions = model_fit.predict(start=len(patent_count), end=len(patent_count) + 10, typ='levels')
```

The model's effectiveness was gauged using metrics like the Mean Squared Error (MSE), ensuring that the predictions are not only insightful but also reliable.

By implementing this method, we obtained a clear picture of how patent trends have evolved over time and where they might be headed in the near future, offering invaluable intelligence for strategic planning and decision-making.

7.4 Rationale

The methods and techniques employed in this project were meticulously selected to harness the intricate depth and breadth of the patent data. The rationale for our choices is anchored in several core principles:

- **Holistic Understanding:** By integrating various data science methods, from sentiment analysis to topic modeling, we aimed to achieve a comprehensive understanding of the data, capturing not just the overt content but also the latent themes and sentiments.
- **Predictive Insight:** Understanding the past and the present is pivotal for forecasting future trends. Our methods, especially topic modeling, offer insights into recurrent and emerging themes, paving the way for predictive analysis.
- **Global Contextualization:** Decoding this patent data isn't merely an exercise in understanding technology in isolation. By analyzing the data, we can position Australia, and more pointedly, Sydney, within the broader global technological landscape. This context is invaluable for NSW-DPC as it navigates strategic planning and policy-making.
- **Maximized Value Extraction:** Every step in our pipeline, from data extraction to transformation and analysis, was designed to extract maximum value. This ensures that not a byte of data is wasted and every piece of information is leveraged to its fullest potential.

In essence, our methods are not just about understanding the data, but about translating that understanding into actionable insights that can drive decision-making for the NSW-DPC. Through this approach, we aim to offer a clear, comprehensive, and context-rich perspective on Sydney's technological trajectory.

8. Findings and Values

8.1 Insights into Technological Trajectories

Our methodical analysis of the patent dataset, which involves techniques like Latent Dirichlet Allocation (LDA) and clustering, aims to unveil the technological trajectories in NSW and Australia. Early findings suggest:

- **Emerging Technology Domains:** By leveraging LDA for topic modelling on patent descriptions and abstracts, we are positioned to identify key technological domains where NSW and Australia are making strides or might need more focus.
- **Temporal Trends:** Our time series analysis aims to offer a chronological perspective on patent publications, revealing the pace at which innovation is growing locally versus international benchmarks.
- **Key Players:** A deep dive into patent holders and classifications will allow us to pinpoint the major corporations, research institutions, and individual innovators central to NSW's innovation landscape.

8.2 International Competitiveness

Utilizing the patent data, our analysis is geared towards understanding NSW and Australia's stance in the global tech innovation space:

- **Comparative Metrics:** Through clustering and topic modeling, we will be able to draw comparisons between the innovation themes prevalent in NSW and those dominating the global landscape.
- **Strategic Insights:** The patterns derived from our analysis will guide policy-makers in understanding where NSW stands vis-à-vis global tech hubs and which areas need strategic intervention.

8.3 Visualization through Power BI Dashboard

Our commitment to data visualization is evident in the interactive displays generated during our analyses. While Power BI will encapsulate the broader patterns, preliminary visualizations:

- **Decision Support:** Offer a nuanced view of patent trends, clusters, and dominant topics, aiding swift and informed decision-making.
- **Accessibility:** The visual narratives derived from complex data ensure that insights are accessible to stakeholders across NSW-DPC, irrespective of their technical background.

8.4 Value Proposition

Our structured and meticulous approach, spanning from the initial stages of NLP data preprocessing (encompassing tasks such as data cleaning, stemming, and lemmatization) to advanced analytical methodologies like topic modelling and time series forecasting, assures the following core values:

Strengthen Strategic Positioning: The rapidly changing technological landscape demands agile, informed decision-making. By unearthing and interpreting patent trends, we are positioning NSW-DPC at the forefront of innovation. Armed with this actionable intelligence, the organization can make strategic moves, predict global shifts, and ensure that NSW remains a trendsetter in the tech world.

Highlight Gaps and Opportunities: Every piece of data tells a story. Our in-depth analyses serve a dual purpose. On one hand, they spotlight areas where NSW is thriving, setting benchmarks in global tech innovation. Simultaneously, they underline domains where there is potential yet to be tapped. By distinguishing these gaps, the NSW-DPC is empowered to steer resources, attention, and initiatives to areas that promise significant growth and can elevate NSW's global technological status.

Promote a Data-Driven Culture: In an era dominated by digital transformation, a data-centric approach is no longer a luxury but a necessity. Our meticulously crafted, user-friendly visualizations are not just tools for interpretation but also instruments of cultural shift. They drive home the message of data's paramountcy. By embedding these tools within the NSW-DPC's decision-making framework, we are fostering a culture that respects, understands, and acts on data-driven insights. This not only augments the decision quality but also ensures that the organization's moves are always backed by empirical evidence, reducing risks and amplifying outcomes.

Empowerment Through Knowledge: Beyond just providing insights, our endeavor is to ensure that every stakeholder, irrespective of their tech proficiency, gains a profound understanding of the data narratives. By integrating easy-to-grasp visualizations with deep-dive analyses, we are ensuring that knowledge isn't just confined to experts. This democratization of information boosts collective intelligence, fosters collaboration, and ensures that all strategies crafted are holistic and representative of the collective wisdom.

9. Challenges & Solutions

9.1 Constructing the Data Pipeline

Challenge: Our maiden venture with Azure services such as Databricks and ML studios presented initial hurdles. Primarily, the limited scope provided by the free account subscription was a significant bottleneck.

Solution: To mitigate these challenges, we streamlined our data processing tasks, optimizing for efficiency. By developing smaller prototype models and utilizing batch processing, we ensured maximum throughput within the given computational constraints. Recognizing that NSW-DPC boasts a professional/business Azure subscription, it was crucial to create a scalable pipeline, allowing the team to harness a more expansive suite of tools and compute power in subsequent iterations.

9.2 Navigating Diverse Data Sources

Challenge: Integrating data from multiple sources often presents inconsistencies. Variabilities in data structures, granularity, and format necessitated a robust data preparation phase.

Solution: By building custom parsing and transformation scripts, we harmonized disparate data sources. Automated validation checks ensured data quality, allowing for seamless integration into our intelligence engine.

9.3 Facilitating Non-Technical Accessibility

Challenge: Translating complex data into a format palatable for non-technical stakeholders is always a daunting task. It's essential to strike a balance between depth and accessibility.

Solution: Our choice to employ Power BI as the visualization tool was pivotal. By leveraging its intuitive design capabilities, we've crafted an interactive dashboard that allows users to dive deep into the insights while ensuring ease of navigation and comprehension.

10. Recommendations & Future Strategy

10.1 Enhancing the Intelligence Engine

As the technological landscape evolves, so will the nature and volume of patents. It's crucial for the strategic intelligence engine to be adaptive and scalable. Our recommendation is to:

- **Incorporate more Data Sources:** Beyond the LENS API, integrating additional databases will offer a more holistic understanding of the global patent landscape. This might include databases from specific regions or industries.
- **Employ Advanced Analytical Techniques:** As we progress, leveraging more sophisticated machine learning and AI techniques will refine the insights extracted. Techniques like Natural Language Processing (NLP) can assist in deeper analysis of patent descriptions, claims, and abstracts.

10.2 Expanding the Power BI Dashboard

The dashboard, while robust, can be further enriched:

- **Custom Alerts and Notifications:** Integrating a feature that notifies stakeholders of significant shifts in patent trends can ensure NSW stays ahead of the curve.
- **User-Customized Views:** Allowing users to save customized views or generate ad-hoc reports directly from the dashboard can significantly enhance user experience.

10.3 Leverage Enhanced Azure Capabilities

Given the superior capabilities of the professional/business Azure subscription available to NSW-DPC:

- **Adopt Bigger Data Models:** Process larger datasets to uncover even more granular insights into global patent trends.
- **Integrate Advanced Azure Services:** Services like Azure Cognitive Services can provide advanced analytics and AI capabilities, refining data analysis.

10.4 Continuous Feedback Loop

Establish a mechanism for continuous feedback from end-users. This will ensure the engine and dashboard remain aligned with the evolving needs of the stakeholders and can be adapted promptly.

11. Conclusion

In retrospect, the Strategic Intelligence Gathering Engine Project underscores the critical importance of adaptive technological frameworks in today's fast-paced digital era. Our collaboration with NSW-DPC set forth an ambitious objective: to harness the power of data and provide actionable intelligence for strategic decision-making.

Throughout the project's lifecycle, we maintained a methodical approach, emphasizing comprehensive data ingestion, intricate analytical processes, and streamlined automation. Our reliance on machine learning models proved pivotal, especially when dissecting vast text data to extract meaningful insights. The implementation of an interactive PowerBI dashboard serves as a testament to our commitment to user-centric design, ensuring that the data is not only understood but can be acted upon with confidence.

Several challenges emerged during the project's execution, particularly when navigating the intricacies of advanced tools such as Azure's ecosystem, Databricks and PowerBI. However, our team's resilience, combined with collaborative problem-solving, ensured that we overcame these obstacles, leaving us with a robust and scalable solution.

The deliverable we present isn't just an end-product but a foundational tool designed for growth and adaptability. Its true potential will be realized when integrated with NSW-DPC's expansive resources and expertise, fostering further refinement and ensuring the platform remains relevant in the face of evolving technological challenges.

As we conclude, it's pertinent to reflect on the broader implications of our endeavor. In an age where data is abundant, the real competitive advantage lies in its effective interpretation. The Strategic Intelligence Gathering Engine stands as a beacon in this direction, offering NSW an edge in the global technological arena. We are confident that our collaboration with NSW-DPC will bear fruit, catalyzing strategic initiatives and reinforcing NSW's reputation as a technological leader.

12. Appendices

Github: <https://github.com/justinqj/iLab>

Notion: <https://www.notion.so/DPC-NSW-Project-Team-4-9d4b3ba3321d41659d18e11ae9f8f937?pvs=4>

Power BI screenshots:

| | | | | | | | |
|------|----------------|---------------|------------|------------|-------------------|--------------|---------|
| Home | Topic Scanning | Date Scanning | Domain Viz | Region Viz | Aus vs. The World | Key Insights | Summary |
|------|----------------|---------------|------------|------------|-------------------|--------------|---------|

NSW TREND INSIGHTS

| | | |
|--|---|--|
| Scanning | Viz | Summary |
| Scanning the journals based on topic, date, and origin | Visualisations of the journals' information | Gaining Insights from the curated journal contents |



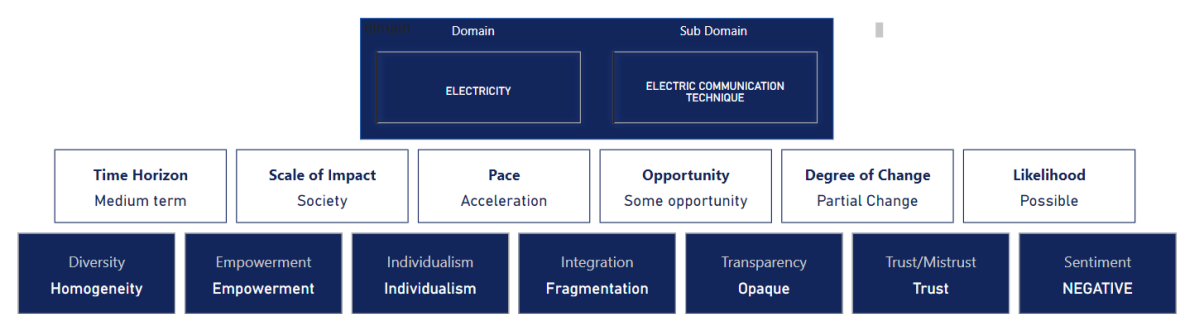
| Topic Scanning | | | | | Home | Topic Scanning | Date Scanning |
|----------------|---------------------|--|-------------------------------------|---|------|----------------|---------------|
| Main Domain | lens_id | invention_title | main_cpc_description | sub_cpc_description | | | |
| All | 090-078-330-939-569 | TRANSMISSION TIME INTERVAL INTEGRATION FOR MULTIPLE RADIO ACCESS TECHNOLOGIES | PERFORMING OPERATIONS; TRANSPORTING | WORKING OR PRESERVING WOOD OR SIMILAR MATERIAL; NAILING OR STAPLING MACHINES IN GENERAL | | | |
| Sub Domain | 172-782-577-350-395 | A MEDICAL SYSTEM AND A DEVICE BASED ON MICROWAVE TECHNOLOGY FOR PREVENTION AND DIAGNOSIS OF DISEASES | PERFORMING OPERATIONS; TRANSPORTING | WORKING OR PRESERVING WOOD OR SIMILAR MATERIAL; NAILING OR STAPLING MACHINES IN GENERAL | | | |
| Jurisdiction | 001-674-572-281-993 | POSITIONNEMENT BASÉ SUR DES SIGNAUX NON CELLULAIRES DE TÉLÉMÉTRIE ET DES SIGNAUX CELLULAIRES DE TECHNOLOGIE D'ACCÈS RADIO (RAT) | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| Search | 008-612-647-650-476 | VERFAHREN, VORRICHTUNG UND BENUTZERENDGERÄT ZUR DRAHTLOSEN KOMMUNIKATION MIT EINER BASISSTATION KLEINER REICHWEITE MIT SCHNITTSTELLEN FÜR MEHRERE FUNKTECHNOLOGIEN | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 010-623-003-502-350 | SYSTEM UND VERFAHREN DER ANGEWANDTEN RADIALTECHNOLOGIE-CHROMATOGRAPHIE | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 012-426-768-400-823 | ZELLABTASTTECHNOLOGIEN UND VERFAHREN ZUR VERWENDUNG DAVON | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 012-949-103-196-322 | SUB-PLATOONS WITHIN VEHICLE-TO-EVERYTHING TECHNOLOGY | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 013-048-430-430-399 | Technologies for adaptive collaborative optimization of internet-of-things systems | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 014-374-306-851-081 | SYNCHRONIZATION IN A PSSS RADIO COMMUNICATION TECHNOLOGY FOR HIGH DATA RATES | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |
| | 016-549-097-795-574 | CODES A BARRES BIDIMENSIONNELS POUR TECHNIQUES DE REPRODUCTION ASSISTÉE | PERFORMING OPERATIONS; TRANSPORTING | WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL | | | |



| Date Published | lens_id | invention_title | Year | Quarter | Month | Day | Year | Quarter | Month | Day |
|----------------|---------------------|--|------|---------|----------|-----|------|---------|-------|-----|
| 11/20/2022 | 080-897-186-328-996 | 360° ASSISTANCE FOR QCS SCANNER WITH MIXE D REALITY AND MACHINE LEARNING TECHNOLOGY | 2023 | Qtr 1 | March | 9 | | | | |
| 8/24/2023 | 073-296-079-369-410 | A BLOCK-CHAIN ENABLED REFERENCE ARCHITECTURE SYSTEM FOR A SMART LEARNING TECHNOLOGY | 2022 | Qtr 4 | November | 24 | | | | |
| | 099-796-184-276-560 | A book production technology | 2023 | Qtr 3 | July | 6 | | | | |
| | 086-212-039-709-637 | A FABRIC FOR A PAPER OR PULP TECHNOLOGY AND A METHOD FOR MANUFACTURING A FABRIC FOR A PAPER OR PULP TECHNOLOGY | 2023 | Qtr 3 | August | 2 | | | | |
| | 135-050-624-409-21X | A FABRIC FOR A PAPER OR PULP TECHNOLOGY AND A METHOD FOR MANUFACTURING A FABRIC FOR A PAPER OR PULP TECHNOLOGY | 2023 | Qtr 3 | August | 10 | | | | |
| | 054-859-429-760-055 | A finish machining technology for manufacturing automation of pump impellers | 2023 | Qtr 3 | August | 10 | | | | |
| | 172-782-577-350-395 | A MEDICAL SYSTEM AND A DEVICE BASED ON MICROWAVE TECHNOLOGY FOR PREVENTION AND DIAGNOSIS OF DISEASES | 2023 | Qtr 3 | August | 10 | | | | |
| | 166-358-346-735-263 | A Method, Network Functions and a Computer Program Product for Supporting the Handing Over of a User Equipment, UE, from a First Type of Radio Access Technology, RAT, to a Second Type of RAT | 2023 | Qtr 3 | August | 17 | | | | |
| | 099-237-985-825-194 | A NEURAL NETWORK FOR IDENTIFYING RADIO TECHNOLOGIES | 2023 | Qtr 3 | July | 27 | | | | |
| | 002-994-605-007-130 | A PROTEIN-INDUCED PLURIPOTENT CELL TECHNOLOGY USES THEREOF | 2023 | Qtr 2 | June | 29 | | | | |
| | 091-767-274-963-686 | A SYSTEM AND A METHOD FOR LIVE STREAMING BY USE OF AN AUGMENTED REALITY (AR) TECHNOLOGY | 2023 | Qtr 1 | February | 16 | | | | |
| | 001-362-826-307-370 | A SYSTEM AND A METHOD TO AUGMENT DEVELOPMENT AND INFORMATION TECHNOLOGY OPERATIONS (DEVOPS) | 2022 | Qtr 4 | December | 14 | | | | |
| | 137-925-191-671-168 | A SYSTEM FOR DEVELOPING AN IOT BASED HEALTHCARE INFORMATION TECHNOLOGY AND A METHOD THEREOF | 2023 | Qtr 1 | January | 19 | | | | |
| | 093-851-928-810-692 | A TIBETAN PIG BREEDING EQUIPMENT AND METHOD BASED ON CIRCULATION COOLING TECHNOLOGY | 2023 | Qtr 1 | March | 30 | | | | |
| | 135-211-357-018-357 | A variable reluctance measurement technology tendon tension monitoring system for mounting on a porch | 2023 | Qtr 2 | April | 20 | | | | |
| | 168-119-494-508-622 | ACCESS CHECK OPTIMIZATION FOR HANDOVER FROM WIRELESS LOCAL AREA NETWORK (WLAN) AND OTHER INTER RADIO ACCESS TECHNOLOGY (IRAT) | 2023 | Qtr 2 | June | 29 | | | | |


000-012-609-013-177

Method, apparatus, server, and systems of time-reversal technology



summary

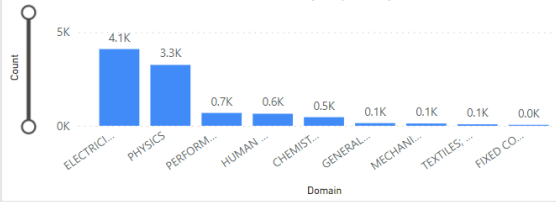
methodological development of SMART is accomplished in two steps. the tool can assist in high-throughput natural product discovery. several newly isolated compounds were automatically located with their known analogues.



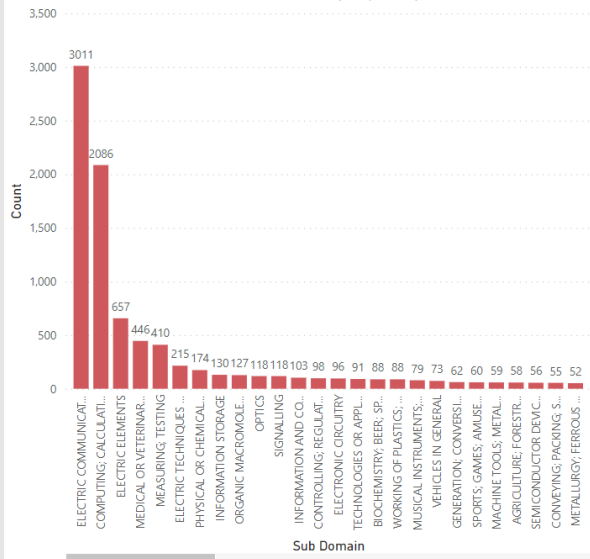
Domain Viz

[Home](#)
[Domain Viz](#)
[Region Viz](#)
[Aus vs. The World](#)

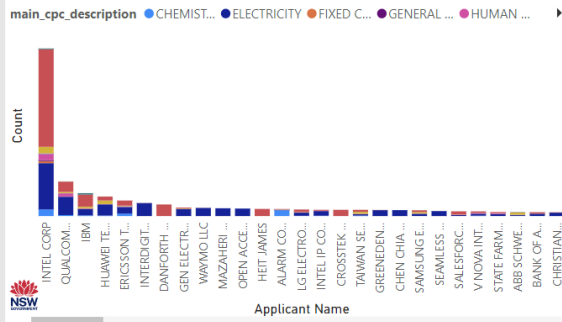
Count of Domain by Popularity



Count of Sub Domain by Popularity



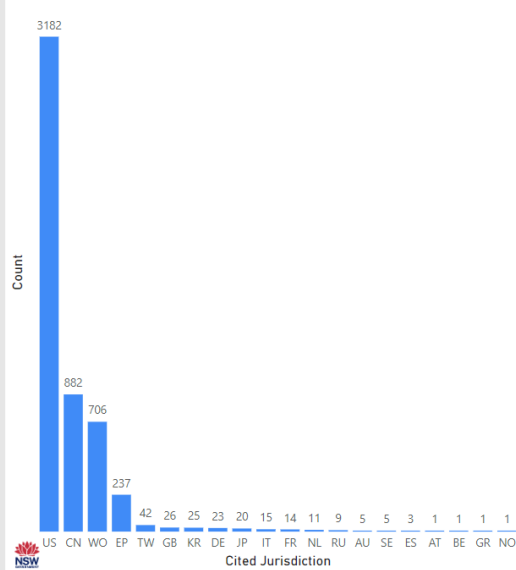
Count by Applicant Name and Domain



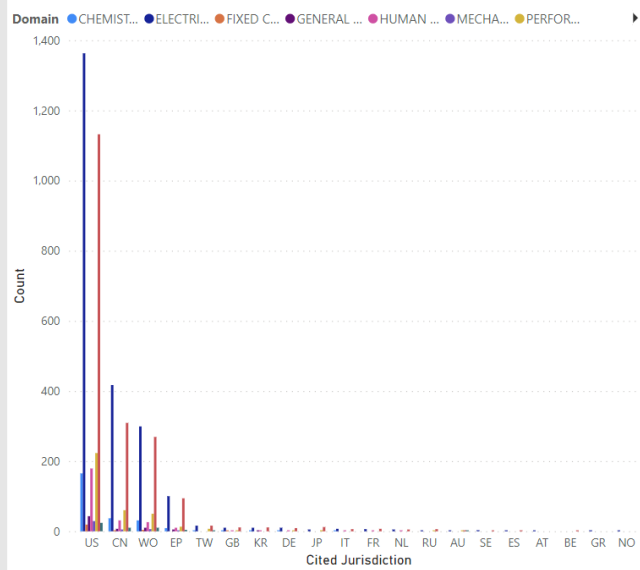
Region Viz

[Home](#)
[Domain Viz](#)
[Region Viz](#)
[Aus vs. The World](#)

Count by Cited Jurisdiction

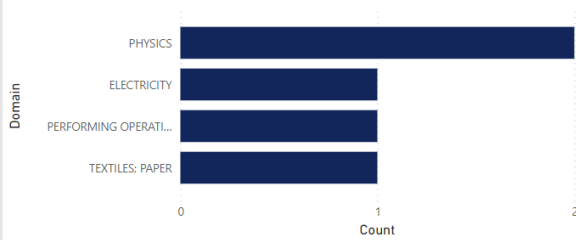


Count by Cited Jurisdiction and Domain

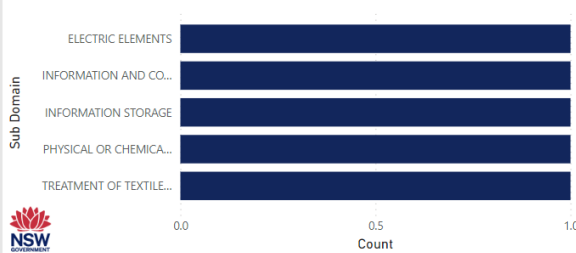


Australia

Count by Domain

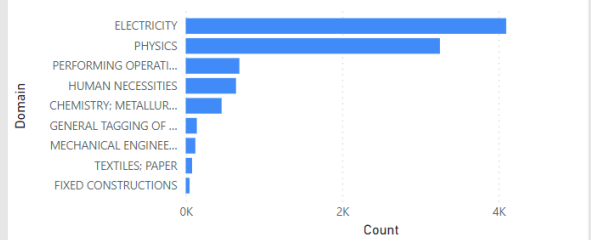


Count by Sub Domain

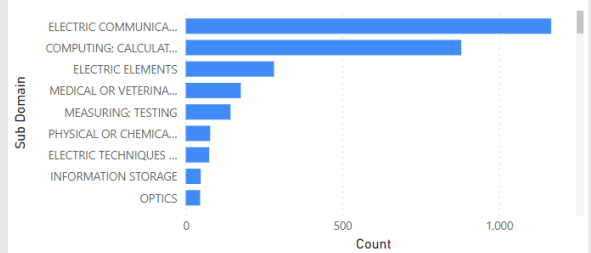


The World

Count by Domain



Count by Sub Domain

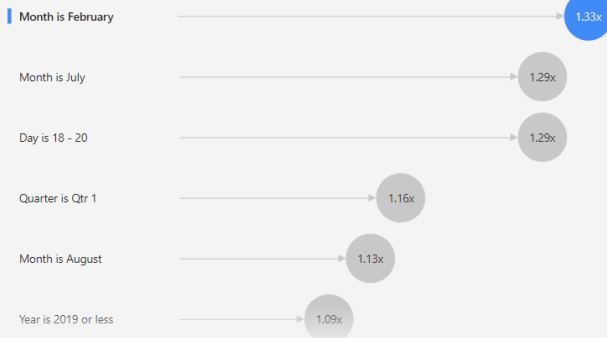


Key influencers Top segments

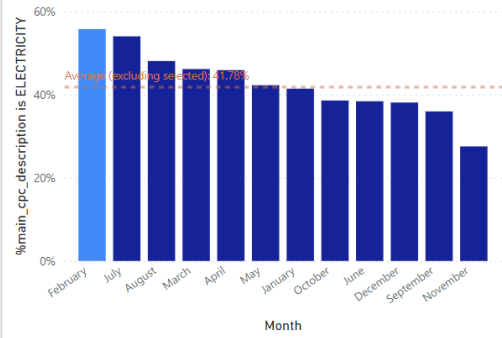
What influences main_cpc_description to be ELECTRICITY ?

When...

...the likelihood of main_cpc_description being ELECTRICITY increases by



← main_cpc_description is more likely to be ELECTRICITY when Month is February than otherwise (on average).

☐ Only show values that are influencers

Key influencers

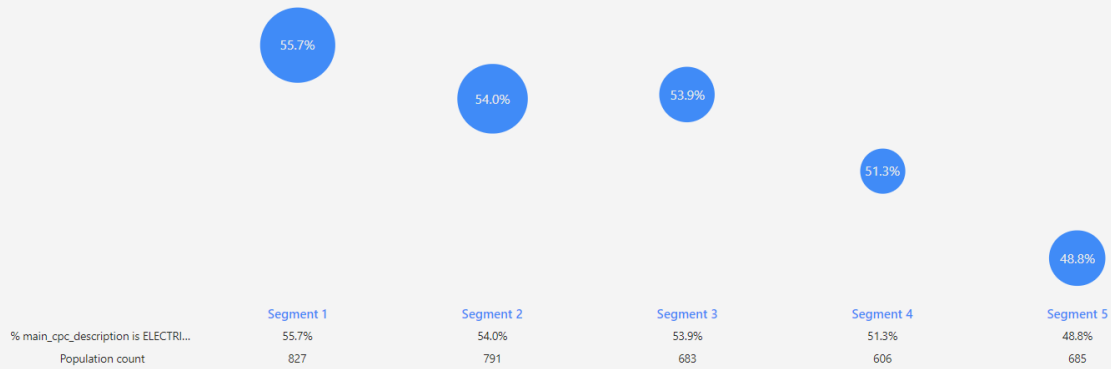
Top segments

When is main_cpc_description more likely to be

ELECTRICITY

 ?

We found 5 segments and ranked them by % main_cpc_description is ELECTRICITY and population size. Select a segment to see more details.



Key influencers

Top segments

When is main_cpc_description more likely to be

ELECTRICITY

 ?

