Using Large Language Models to Automate Patent Citation Analysis

CAPSTONE PROJECT REPORT SUBMITTED

TO THE

INDIAN SCHOOL OF BUSINESS

FOR THE

ADVANCED MANAGEMENT PROGRAMME IN BUSINESS ANALYTICS

SUBMITTED BY:

No.	Student Name	PGID	AMPBA Email ID
1	Archit Goel	12220026	Archit_Goel_ampba2023W@isb.edu
2	Arpit Agarwal	12220065	arpit_agarwal_ampba2023w@isb.edu
3	Dhruv Puri	12220058	Dhruv_Puri_ampba2023W@isb.edu
4	Kunwar Ji Gupta	12220025	kunwar_gupta_ampba2023w@isb.edu
5	Paritosh Sinha	12220056	Paritosh_sinha_ampba2023w@isb.edu

UNDER THE GUIDANCE OF

Mr. Bharani Kumar

INSTITUTE OF DATA SCIENCE

INDIAN SCHOOL OF BUSINESS

MONTH YEAR OF SUBMISSION OF REPORT

May 2024

CAPSTONE PROJECT REPORT

SUBMITTED TO THE INDIAN SCHOOL OF BUSINESS FOR ADVANCED MANAGEMENT PROGRAMME IN BUSINESS ANALYTICS

HEADING	**DETAILS**
Using Large Language Models to Automate Patent Citation Analysis	Citations are important component in the field of Patents. Citations are integral part of various activities in the world of Patent such as patent application, search, patent infringement etc. Searching through patent world for above objectives is hugely time and cost intensive.

	Our project aims to improve accuracy of patent search over existing attempts made to automate Forward & Backward citations using LLM and its variants.
Name of the Candidate	**Mr. Archit Goel** **Mr. Arpit Agarwal** **Mr. Dhruv Puri** **Mr. Kunwar Ji Gupta** **Mr. Paritosh Sinha**
Name and Affiliation of the Faculty Mentor	**Mr. Bharani Kumar; Faculty,** **Indian Institute of Business**
Name and Affiliation of the Industry Mentor	**Mr. Prem Kalyan &** **Mr. Radhakrishnan** **(Accolite Digital)**
Batch	**AMPBA 2023(Winter)**
Date/Month of submission of Capstone Project Report	**May 2024**

TABLE OF CONTENTS

Chapter No.	Topics
	List of Tables
	List of Figures
	Abstract
	Executive Summary
Chapter 1	**Introduction**
1.1	Intellectual Property Right (IPR)
1.2.1	Types of IPR
1.3	Patents
1.3.1	Types of Patents
1.3.2	Criteria for Patentability
1.3.3	Non-Patentable Subject Matter
1.3.4	Parts of a patent Application
1.3.5	Patent Family
1.4	Citations
1.5	Granting of Patents
1.5.1	Important Dates in a Patent Application
Chapter 2	**Modular Description of the Job**
2.1	Categories of Patent Applications
2.2	Claims
2.2.1	Categories of Claims

Chapter 3	**Patent Searching**
3.1	Patentability Search/ Novelty search
3.2	Validity Search
3.3	Infringement Search
3.4	Freedom-To-Operate Search
3.5	State of the Art Search

3.6	Business/Research Problem
3.7	Data Available
Chapter 4	**Assumptions, Approach & Process**
4.1	Assumptions
4.2	Approach
4.3	The Hugging Face Framework
Chapter 5	Exploratory Data Analysis
5.1	Features/ Notable findings
Chapter 6	Feature Selection and Engineering
6.1	Introduction
6.2	Chunking
6.3	Pros and Cons
6.4	Text Summarization
6.5	Embeddings
6.6	Feature Integration
6.7	Challenges and Solutions
6.8	Conclusion
Chapter 7	Models Used
7.1	Models and Metrics Considered
7.2	Models Chosen
7.3	Results of the models
7.4	Model Deployment and Evaluation
Chapter 8	Conclusion /Recommendations
Chapter 9	References

List of Tables

Table No.	Title	Page No.
5.1	Title of the Table	22
7.1	Model Description	32
7.2	Model Assumptions	33

List of Figures

Figure No.	Title	Page No.
1.1	Different forms of IPR	9
1.2	General Characteristics of Utility Patents	10
5.1	CRISP-ML(Q) Architecture	20
5.2	Project Architecture	23
6.1	Feature Section	28
6.2	Text cleaning using RegEx	30
7.1	Chunk creation	33
7.2	Text summarization with other RAG model (RAG tokenizer)	36
7.3	Text summarization with Pegasus RAG mode	37
7.4	Text before embeddings	40
7.5	Text converted to embeddings	40
7.6	Final recommendation list	44

Abstract

In the information space of intellectual property rights (IPR) and innovation, the reality is traditional patent search methods have not performed effectively and could not deliver and fulfill requirements. Not only can we achieve that by utilizing the latest technology in the language models and search methods, but we developed an improvement to handle these challenges and victoriously modernize patent analysis and recommendation processes.

Adopting the most recent Retrieval-Augmented Generation (RAG) approaches for response generation, we have tried to build our patent analysis framework which is novel in terms of methods. Integrated cascading metadata with document content and incorporating multi rooted retrieval ideas, the aim of the framework is to establish accuracy and specific searches in patent searches.

Our RAG framework, which builds on the existing systems in healthcare and software engineering, aims to achieve the same query-intent detection accuracy that is crucial in patent research. To reach this objective, we are linking external databases and using advanced language models. This way, we believe

that our framework can be more trustable and precise than just one person's knowledge, that may be not updated, and the understanding processes which are arbitrary.

Using total data acquisition, advanced text data prepping, and frameworks modeling, it probes to simplify the patent search process while maintaining its better recommendations reliability and credibility. We intend to fix the problems as stated by some researchers, which are the limitations in the retrieval systems is the main one and we will only rely on the models to create a robust and scalable solution for patent analysis and recommendation.

In short, our framework bridges the gap between traditional patent search methods and the evolving needs of innovators and organizations. By harnessing the power of advanced LLMs and retrieval methodologies, our framework promises to empower stakeholders with accurate, timely, and actionable insights.

Key words:

1. Patent search
2. Forward and Backward Citations
3. Retrieval Augmented Generation (RAG)
4. Text compression and summarization
5. Chunking & Indexing
6. Vector Embeddings and vector store
7. Elastic Search, Lexical search, Sparse (for instance, BM25) and Dense search
8. Ranking and recommendation algorithms and systems

Executive Summary

In today's dynamic business world Innovation helps you remain competitive. But to innovate One requires constant confluence of Ideas and designs.

However, to encourage innovation in ideas and design. One requires protection. Not only for continuation and competitiveness of business but also to support continuous innovation in ideas and design. Patenting Ideas and designs is one way of protecting the rights of innovators.

However, getting Innovation Patented has not been an easy job. Because traditional patent searching methods often present challenges such as inefficient search strategies, Inaccurate results, time and resource constraints and legal complexities. These limitations hinder the effectiveness and efficiency of patent search impacting innovation. Market competitiveness, legal compliance and economic growth.

To address such challenges, we require innovative solutions. Such solutions can be built around technological advancement made in the field of machine learning through LLMs and RAG models. Such solutions can enhance patent searching capabilities and unlock new opportunities for innovation and growth.

Through LLM and RAG models we can go and adapt strategies which go beyond simplicity, simplistic keyword based approaches.

1. Innovation and technology development: Advanced search tools and technologies facilitate faster and more accurate identification of relevant patents.
2. Market competitiveness and strategic decision making: Improve patent search capabilities enable businesses to make informed decisions regarding product development, market entry strategies, and intellectual property management.
3. Legal compliance and risk management: Accurate patent searches are essential for ensuring legal compliance and mitigating the risk of patent infringement &/or litigation.
4. Knowledge sharing and collaboration.: Standardization and integration of patent databases promote knowledge sharing and collaboration among stakeholders accelerating the pace of innovation.
5. Economic growth and job creation: Enhanced patent search capabilities drive economic growth and job creation by stimulating innovation, Supporting entrepreneurship and attracting investments in R&D.

A model built considering above considerations should have success criteria such as reduction in time consumption, Precision in recommendation generation and economic efficiency.

To conclude enhancing patent search efficiency and effectiveness requires a multifaceted approach that integrates advanced technologies, collaboration and standardization. By embracing innovative solutions and leveraging insights from recent research, businesses can navigate the complexities of the patent landscape with greater ease, driving innovation, competitiveness and economic growth.

Chapter 1: Introduction

1.1 Intellectual Property Rights (IPR)

Intellectual property means products of the mind that individuals can protect as an intangible asset. It has got similar legal protection as other property rights.

These can span domains such as industrial, scientific, literary, and artistic, taking the form of inventions, manuscripts, software suites, or business names.

1.2 Types of IPR

• **Copyright:** Copyright is a legal construct, granting the creator of an original work exclusive rights to its use and distribution for a limited time. This is intended to enable the creator of intellectual wealth to receive compensation for their work. Copyright remains valid for the lifetime of authors, plus 50 years after their death.

- **Patent:** A patent is a legal right granted by a government to an inventor or assignee for a limited period. In this case a detailed public disclosure of an invention is made. These are novel inventions as solutions to specific technological problems, products, or processes. Validity extends upto 20 years.
- **Trademark:** A trademark is a recognizable sign, design, or expression distinguishing products or services of a particular source from those of others.
- **Geographic Indications:** This constitutes an industrial property right on goods with a specific geographical origin, possessing unique qualities or a reputation attributable to that origin. Example: Basmati rice.

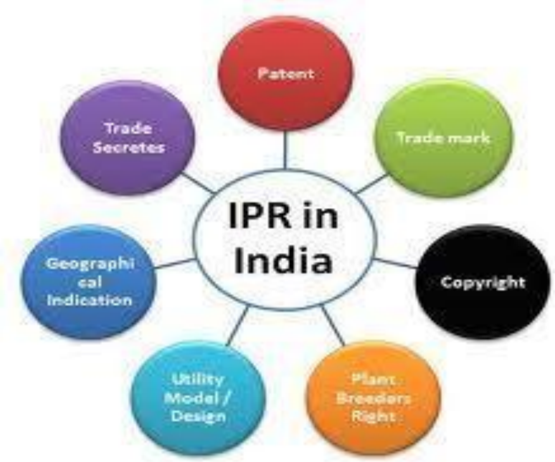


Fig: 1.1 Different forms of IPR

- **Trade secret:** A trade secret could be a formula, practices, processes, etc. not generally known or accessible, through which a business can gain an economic advantage over competitors or customers example COKE or KFC product formulation

1.3 Patents

A patent denotes an exclusive right bestowed upon individuals who devise any novel, beneficial, and not readily apparent process, machine, article of manufacture, or composition of matter, or any innovative improvement thereof. It is granted to inventions that offer fresh methods of operation or enhanced solutions to technical dilemmas. Upon meeting the criteria of patentability, an invention can be safeguarded as a patent, valid for 20 years from the date of filing the patent application.

Advantages:

The patent system can be viewed as an encouragement or incentive for individuals to be creative, since the effort and hard work is in that individuals who create the innovations are recognized and rewarded. Hence, such a recognition should lead to creativity and in turn, this would consequently raise the standard of life. Besides, patents provide prospects for the patent owners to make a living by charging fees from others who want to use their technology.

Disadvantages:

However, filing for a patent incurs significant costs and entails liability.

1.3.1 Types of Patents

- **Utility Patent:** Within this we will consider something that is indeed new, namely new processes, machines, compositions of matter or even improvement of known ideas. Its scope is within a period of 20 years from the patent application date.

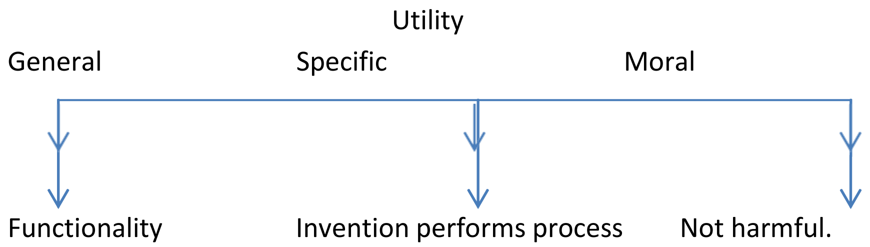


Fig: 1.2 General Characteristics of Utility Patents

- **Design Patent:** It covers aesthetic value and does not possess functional attributes. For example, designs of chairs, wallpapers, shoes, and jewelry fall under this category. Its validity extends for 15 years from the date of grant.
- **Plant Patent:** This covers patents for plants reproduced asexually. Its validity persists for 20 years from the date of filing.

1.3.2 Criteria for Patentability

- **Novelty:** Novelty stands as a most basic requirement for patentability. An invention is deemed novel only if it has not been previously known in any oral or written form. The purpose of this requirement is to prevent prior art from being patented again.

- **Inventive Step and Non-obviousness:** An invention must be sufficiently inventive—i.e., non-obvious—to qualify for patent protection.
- **Utility:** The invention should demonstrate some industrial utility, meeting certain requirements to be useful in an industry.

1.3.3 Non-Patentable Subject Matter

Certain subject matters are ineligible for patent protection, including laws of nature, abstract ideas, mental processes, printed matter, computer software, and methods of doing business.

1.3.4 Parts of a Patent Application

A patent application, as well as a granted patent, comprises various parts, each holding its own significance: A patent application, as well as a granted patent, comprises various parts, each holding its own significance.

- **Title:** This part will be called "The name" of the invention. This will be the protected invention part.
- **Abstract:** Join the introduction, indicate the system that will be patented.
- **Field of Invention:** This deals with the discussion of the invention.
- **Background:** The brief recount revealing the prior or related art regarding the invention, pointing out new technology's limitations and showing how the present invention leaps forward to the present stage.
- **Summary:** Summing up the invention in a few words by juggling all the subtle details in one sentence.

• **Brief Description of Drawings:** Through a concise description of drawings that will be attached, the application will provide an opportunity to display my artwork and represent the cultural foundations of my ethnic background.
• **Detailed Description of Drawings**: Through careful and insightful presentation of each image presented in your application if you so wish.
• **Claims:** These mark out the legal line that divides the patent privileges that are conferred.
• **Drawing:** Graphics serve as an accompanying material to the patent application.

1.3.5 Patent Family

A patent family constitutes a collection of all patent applications and granted patents filed across multiple countries to safeguard a single invention by one or more common inventors.

Initially, a primary application is submitted in one country—known as the priority country—and subsequently extended to other jurisdictions.

An INPADOC (International Patent Documentation Centre) patent family this encompasses all patent documents resulting from an initial filing with a patent office and subsequent filings within the priority year in other countries.

1.4 Citations

Citations in a patent refer to references to earlier prior arts. Some references are provided by the applicant, while others are cited by the examiner who assesses the patent application.

Backward Citations: These pertain to references of prior arts used in a patent, cited by both the applicant and the examiner.

Forward Citations: These references highlight inventions in the same field made after the patent is issued. Patents citing the subject patent as prior art are termed forward citations, facilitating patent searches.

1.5 Granting of Patents

The national patent offices of countries where the applications were submitted conduct evaluation for the applicants or there is the regional office which conducts assessments concerning a group of countries.

Patent applicants thus can apply for protection of the very same invention within different regional systems and at the end of the day a decision is made on whether to be granted protection in each country providing its jurisdictions.

The Hague Agreement of the WIPO, administered by the same institute, would grant you an international patent, working as efficiently as national applications of various countries filed.

1.5.1 Important Dates in a Patent Application

Certain dates bear significant importance in a patent application, defining its lifespan and legal status:

- **Invention date:** Indicates when the invention was completed.
- **Filing date**: Marks the submission date of the application with all required information.
- **Priority date**: Signifies the first filing date of the application anywhere globally.
- **Issue date:** Denotes the grant date when the patent is officially issued by the patent office.
- **Expiration date**: Indicates the termination date of the patent term.
- **Publication date:** Marks the date when patent information becomes publicly available, typically 18 months after the priority date.

Chapter 2: Review of Literature

2.1 Categories of Patent Applications

- **Regular Application:** At the initiation of the patent application, it is filed in the Patent office with the coalition that the applicant is not claiming the priority on any already filed application, or as a normal patent application. This is an incomplete application, which includes only the Provisional specification, with a detailed description of the invention and claims.

Similarly, it may be a Provisional application if it does not provide details of a particular invention, but only provides the bare idea.

A provisional application is advertised only to reserve its filing date and is not exposed to the examination.

After the Provisional application year has elapsed, the remaining steps would be to prepare and file a Non-Provisional application.

- **Convention Application:** A patent application wherein an applicant asserts a priority date based on the same or substantially similar application filed in one or more convention countries is known as a convention application.

To qualify for convention status, an applicant must file the application within 12 months from the date of the first application in the convention country.

- **PCT- International Application:** The World Patent Cooperation Treaty (PCT) simplifies the process of making international claims for patents. These types of forms may, for instance, be used in any states which belong to the PCT to cover the whole variety of countries in the world during the 'National Stage' while you do not need to go and file them separately in all participating countries.

' PCT application does not imply issuance of a unique international one but significantly simplifies obtaining a patent in the target countries that was originally filed in, at the same time.

- **PCT - National Phase Application:** Following the international phase, the PCT-national phase commences. The applicant must individually 'enter into the national phase,' i.e., file a national phase application in each country they wish to pursue. The applicant can enter the national phase in up to 138 countries within 30-31

Failure to enter the national phase within the stipulated time frame leads to the loss of effect of the International Application in the designated or elected States.

- **Continuing Patent Application:** These applications are of 3 types and are filed for patent protection concerning any enhancement or modification of an invention outlined or disclosed in the complete specification of an already submitted application or a granted patent.

2.2 Claiming

Claiming defines boundaries to the extent of protection provided by a patent or the protection sought in a patent application.

It sets the scope of protection bestowed by the patent.

2.2.1 Categories of Claims

- **Independent Claims:** These stand alone and are self-contained, invariably broader than subsequent dependent claims.

• **Dependent Claims:** These rely on a parent claim and refer back to it, allowing inclusion of all limitations from the parent claim.

They help encompass the invention and its various embodiments, narrower in scope than the parent claim. Features may be added but not deleted from the parent claim.

• **Multiple Dependent Claims**: These refer to more than one other claim and must do so alternately. Multiple dependent claims cannot serve as a basis for any other multiple dependent claims and entail higher filing fees.

CHAPTER 3: Traditional Techniques of Patent Citation Analysis

Patent citation analysis involves systematically identifying and evaluating prior work cited within patent documents to understand the innovation landscape and the influence of patents. This process leverages the well-organized and consistent nature of patent literature, allowing for highly systematic methodologies.

Patent analysts use various tools and techniques, including Citation Searching, Bibliographic Data Searching, Classification Searching, and Full-Text Searching, to conduct comprehensive analyses. Below are the types of searches and their relevance to citation analysis.

3.1 Patentability Search/ Novelty search

Patentability searches help inventors determine if prior art exists that could impact the novelty and patentability of their invention. These searches aim to identify preceding art and assess novelty and non-obviousness.

Key Considerations:

- Time Constraints: Typically require 4 to 20 hours of work.
- Alternative Embodiments: Searches may reveal documents showing different implementations of the invention, which can be crucial for comprehensive novelty assessments.
- Major Patent Collections: Extensive searches in major collections (US, EP, WO/PCT, JP) to ensure thorough coverage.
- Tool Selection: Optimal tools balance completeness with budget constraints, especially for shorter searches.

3.2 Validity Search

Validity searches evaluate whether the prior art relevant to the claims being examined affects their novelty and non-obviousness. This involves a detailed analysis of each claim.

Patents can be invalidated if earlier inventions are discovered. Prior art searches also determine the value of a patent application, ensuring no similar inventions or public disclosures invalidate the patent.

3.3 Infringement Search

Infringement searches identify products that infringe on a patent's claims by being sold without the patent holder's permission. These searches focus on identifying post-grant products infringing on the patent claims.

Key Points:

- **Territorial Specificity:** Infringement can only occur in countries where the patent is in force.
- **Sources:** Primarily patent databases, but may include non-patent sources like product literature.

3.4 Freedom-To-Operate Search

This search examines issued or pending patents to determine if a new product infringes any existing claims. It also considers expired patents that might allow usage of the product despite existing patents. Freedom-to-operate analysis ensures that commercializing a product does not violate others' intellectual property rights in specific jurisdictions.

3.5 State of the Art Search

State of the art searches provide an overview of current developments in a specific field, helping guide research and development. These searches involve analyzing patents related to specific technologies to avoid infringing existing patents and to identify emerging trends.

CHAPTER 4: Project Description & Overview of Patent Citation Analysis

4.1 Project Overview

The landscape of patent research is ripe for transformation, with traditional methods facing inefficiencies and inaccuracies. Leveraging Large Language Models (LLMs), this project aims to automate patent citation analysis, enhancing efficiency, accuracy, and accessibility while providing valuable insights into technological trends and innovations.

The project seeks to revolutionize patent research by harnessing the power of LLMs to automate citation analysis, thereby addressing the shortcomings of traditional methods and empowering stakeholders with actionable insights for informed decision-making.

4.2 Business Problem

Traditional Patent Search Methods Suffer From Inefficiencies And Limitations, Leading to Litigations & Losses:

- **Inefficient Strategies:** Keyword-based Searches Overlook Vital Patents And Prior Art
- **Inaccurate Results**: Manual Reviews Lead To Errors And Legal Risks
- **Resource Constraints:** Time, Expertise, And Financial Demands Hinder Innovation
- **Legal Complexity:** Varied Laws Require Specialized Knowledge For Accurate Searches

4.3 Business Objective

- **Enhance Accuracy:** Improve precision and exhaustiveness in patent search.
- **Efficiency:** Utilize LLM-based search engine to save time and resources by quickly pinpointing relevant patents.
- **Improve Efficiency:** Implement strategies to balance thoroughness and speed, ensuring comprehensive results while minimizing resource constraints.

4.4 Business Constraints

- **Expertise Constraints:** Limited access to specialized patent and technical knowledge.
- **Budgetary Limits:** Financial constraints hinder adoption of advanced search tools or hiring expertise.
- **Technological Barriers:** Legacy systems may obstruct integration of advanced search algorithms.

Chapter 5: Assumptions, Approach & Process

5.1 Assumptions

In constructing a patent citation model using RAG (Retrieval Augmented Generation) and LLM (Large Language Models), we operate under several key assumptions to ensure the model's efficacy and relevance:

1. **Data Quality**: We assume that the patent data used for the learning process is accurate, comprehensive, and free from significant errors. High-quality data is critical for training an effective model, and any inaccuracies or omissions could negatively impact the model's performance and the reliability of its outputs.
2. **Relevance of Citations**: It is presumed that the citations included in the patent data are relevant and represent genuine relationships among the patents. This means that the citations are not arbitrary or coincidental but rather reflect meaningful connections that the model can learn and replicate.
3. **Model Generalization**: We expect the trained model to generalize well to new, unseen patent data. This implies that the model can maintain high-quality citation relationships even when applied to different datasets that it was not explicitly trained on.
4. **Legal and Regulatory Considerations**: The model is developed with the understanding that it must comply with all relevant legal and regulatory requirements concerning patent data usage, intellectual property rights, and privacy regulations. Ensuring compliance is essential to avoid legal issues and maintain ethical standards.
5. **Homogeneity of Data**: The assumption is that the data used to train and test the model is homogeneous in terms of patent types, industries, and jurisdictions. This consistency helps in ensuring that the model's learning is not biased by variations in data sources.
6. **Model Interpretability:** The model should produce interpretable and explainable outcomes, allowing stakeholders to understand and verify the citations it recommends. This transparency is crucial for trust and adoption by end-users.
7. **Bias in Data or Model**: We assume that the model, designed for patent citation analysis, will perform effectively across different categories of patents. This includes handling potential biases in the data and ensuring that the model remains fair and unbiased in its recommendations.
8. **Scalability**: The components of the model are expected to be scalable. This means that the methods and technologies used can handle large volumes of data efficiently and can be scaled up as the size of the patent dataset grows.

5.2 Approach & Methodology

CRISP-DM Framework Integration in System Architecture Report

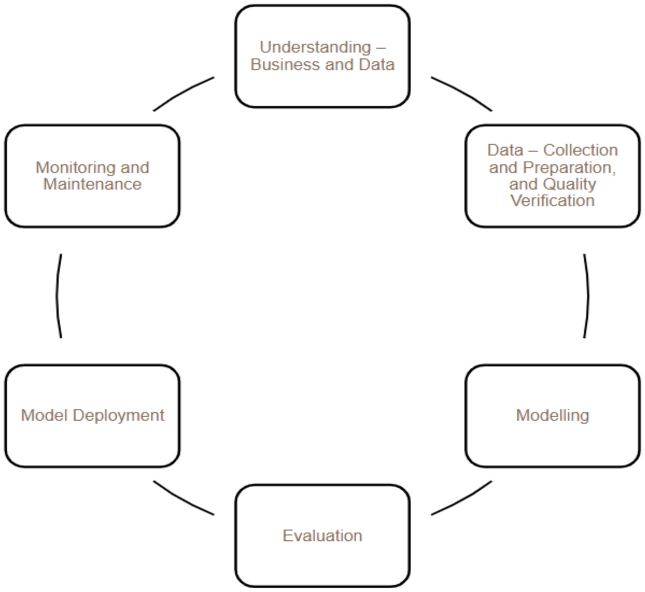

Understanding – Business and Data
Data – Collection and Preparation, and Quality Verification
Modelling
Evaluation
Model Deployment
Monitoring and Maintenance

Fig: 5.1 CRISP-ML(Q) Architecture

➔ **Understanding – Business and Data**

- **Understanding Business Objectives**: Collaboratively understanding the business goals alongside the data requirements allowed us to form insightful queries on the H2 Database.
- **Exploratory Data Analysis (EDA)**: Extracting key features crucial for subsequent model tasks through exploratory analysis. This step enhanced data value by transforming raw text into an informative feature set, crucial for accurate Large Language Model (LLM) tasks in patent citation analysis.

➔ **Data – Collection and Preparation, and Quality Verification**

- **Data Collection:** Utilizing the H2 Database as the primary data source, we meticulously collected patent-related data to ensure completeness and relevance.
- **Data Preparation:** , the pre-processing stage involved
  - **Text-Based Data Cleaning**: Employing NLP libraries for text normalization, spell checking, and punctuation removal.
  - **EDA and Feature Engineering**: Extracting key features crucial for subsequent model tasks through exploratory analysis. This step enhanced data value by transforming raw text into an informative feature set, crucial for accurate Large Language Model (LLM) tasks in patent citation analysis.

- **Quality Verification:** Thorough validation checks were performed at each stage to maintain data integrity and consistency.

➔ **Modeling**

- **Chunking:** Breaking down text into manageable chunks without losing semantic meaning, ensuring efficient processing while retaining contextual relevance.
- **Summarization using RAG:** Generating concise summaries of patent-related information using Retrieval-Augmented Generation (RAG), which combines retrieval and generation capabilities to produce meaningful summaries.
- **LMM-based Embedding Generation:** Leveraging a modified and fine tuned Pegasus Transformers model to create embeddings that capture the semantic essence of the text. These embeddings are crucial for understanding underlying patterns and relationships in the data.
- **Embedding Storage and Analysis:** Storing the generated embeddings in a vector database, such as Faiss, optimized for efficient similarity search and clustering operations. Clustering and similarity analysis are then performed to identify and group similar patents, facilitating accurate recommendations.

➔ **Evaluation**

- **Model Performance Evaluation:** Assessing the effectiveness of the model in generating recommendations based on predefined metrics such as accuracy, precision, and recall using **BLEU Score, Rogue Score** and **BERT Score**.
- **Feedback Incorporation:** Incorporating user feedback into the evaluation process to iteratively enhance model performance.

➔ **Model Deployment**

- **System Integration:** Deploying the model on an EC2 instance, we created a Python script that takes user input and outputs potential patents which could be treated as citations in the form of recommendations. This ensures smooth deployment and operation without adding unnecessary complexity.
- **User Interface Design:** Designing a simple yet effective user interface that allows users to easily input data and receive recommendations, enhancing user experience and accessibility.

➔ **Monitoring and Maintenance (Future Scope)**

- **Pipeline Monitoring:** Implementing a robust monitoring framework to track system performance, detect anomalies, and identify issues in real-time. This ensures the system operates smoothly and reliably.
- **Regular Maintenance:** Conducting routine maintenance tasks to address system updates, data drift, and performance degradation. This proactive approach ensures continuous operational efficiency and long-term success.

Conclusion

Our system architecture for recommending potential citations for a patent demonstrates a systematic approach to solving real-world challenges. Aligning our methodology with the CRISP-DM framework, our solution integrates advanced data processing techniques, RAG-based summarization techniques along with LLM-based Feature Extraction and Embedding Generation to provide users with valuable recommendations derived from patent information.

5.3 Process & Architecture

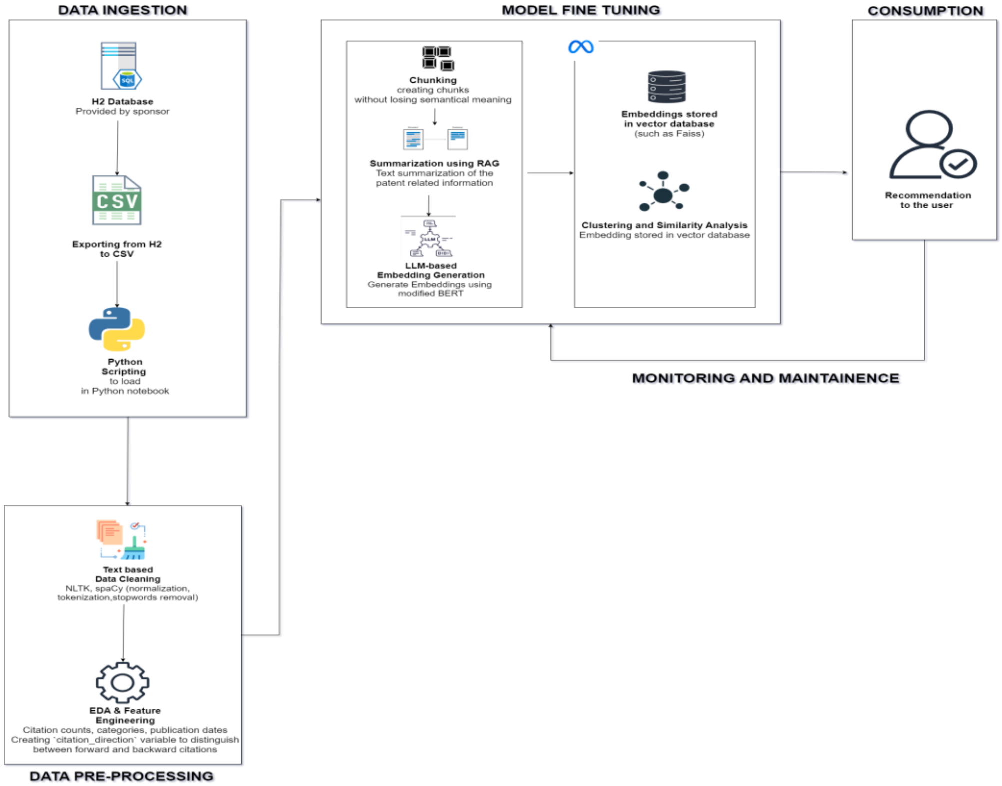

DATA INGESTION
H2 Database
Provided by sponsor
SQL
CSV
Exporting from H2
to CSV
Python
Scripting
to load
in Python notebook
MODEL FINE TUNING
Chunking
creating chunks
without losing semantical meaning
Summarization using RAG
Text summarization of the
patent related information
LLM
LLM-based
Embedding Generation
Generate Embeddings using
modified BERT
Embeddings stored
in vector database
(such as Faiss)
Clustering and Similarity Analysis
Embedding stored in vector database
CONSUMPTION
Recommendation
to the user
MONITORING AND MAINTAINENCE
Text based
Data Cleaning
NLTK, spaCy (normalization,
tokenization,stopwords removal)
EDA & Feature
Engineering
Citation counts, categories, publication dates
Creating `citation_direction` variable to distinguish
between forward and backward citations
DATA PRE-PROCESSING

Fig: 5.2 Project Architecture

Chapter 6: Features/ Notable findings

6.1 Data Available

1. **g_applicant_not_disambiguated**

    a. Data Elements: Contains elements such as patent_id, applicant_sequence, raw_applicant_name_first, raw_applicant_name_last.
    b. Definitions: Includes definitions like the patent number, order of applicants, first name, and last name of the applicant if the applicant is an individual.

2. **g_assignee_not_disambiguated**
    a. Data Elements: Includes elements like patent_id, assignee_sequence, raw_assignee_name.
    b. Definitions: Defines the patent number, the order of the assignees, and the raw assignee name as it appears in the data.
3. **g_inventor_not_disambiguated**
    a. Data Elements: Comprises elements such as patent_id, inventor_sequence, raw_inventor_name_first, raw_inventor_name_last.
    b. Definitions: Defines patent number, order of inventors, first name, and last name of the inventor.
4. **g_uspc_current**
    a. Data Elements: Contains elements like patent_id, mainclass_id, subclass_id.
    b. Definitions: Specifies the patent number, current main class, and subclass identifiers according to the US Patent Classification.
5. **g_cpc_current**
    a. Data Elements: Includes elements such as patent_id, cpc_subgroup_id, cpc_group_id.
    b. Definitions: Defines the patent number, CPC subgroup, and group identifiers.
6. **g_foreign_priority**
    a. Data Elements: Consists of elements like patent_id, foreign_priority_id, priority_date.
    b. Definitions: Includes definitions for patent number, foreign priority number, and the date of the foreign priority.
7. **g_mainclass_current**
    a. Data Elements: Comprises elements such as mainclass_id, mainclass_title.
    b. Definitions: Defines the main class ID and the title associated with it.
8. **g_subclass_current**
    a. Data Elements: Includes elements like mainclass_id, subclass_id, subclass_title.
    b. Definitions: Specifies the main class ID, subclass ID, and the title of the subclass.
9. **g_wipo**
    a. Data Elements: Contains elements like patent_id, wipo_field_sequence, wipo_field_id, wipo_sector_title, wipo_field_title.
    b. Definitions: Defines patent number, sequence order of WIPO technology fields, WIPO field ID, sector title, and field title.

6.2 Notable Features

1. Patent Details: The 'PATENT' table serves as the core, containing crucial details such as patent ID, type, date, title, abstract, and classification details following WIPO standards.

2. Applicant Information: Detailed records of applicants, including names, organizations, types, and associated authorities, offer insights into patent ownership and applicant demographics.

3. Assignee Data: Similar to applicants, assignee data provides valuable insights into entities to whom patent rights are assigned.
4. Inventor Profiles: Inventor details, accompanied by disambiguated names and gender codes, facilitate inventor-based citation analysis.
5. Geographical Information: Both standardized and raw geographic locations associated with patents enable geographic trend analysis in patent filings.
6. Patent Claims and Descriptions: Detailed claims and descriptions offer extensive data for textual analysis, revealing the scope and breadth of the patents.
7. Citation Analysis: Tables specifically dedicated to U.S. application and patent citations enhance the dataset's capability to trace the influence and lineage of patents, essential for citation analysis.

6.3 Feature Selection

Objective of Feature Engineering: The purpose of feature engineering in the current project is to transform non-structured raw patent text data into an elaborated and informative feature set that improves accuracy of the Large Language Model (LLM) in performing tasks of patent citation analysis.
Overview of Techniques: We applied the methods like chunking, text summarization, and embeddings to preprocess and provide the engineered features to the patent data.
1. The dataset comprises 13 primary tables with varying dimensions, each structured with multiple fields encompassing character varying, numeric, date, and text objects.

- Applicant_Raw
- Assignee
- Inventor
- Location
- Location_Raw
- Patent
- Patent_Claims
- Patent_Detail_Desc
- Patent_Draw_Desc
- Patent_Foreign_Citation
- Patent_US_Application_Citation
- Patent_US_Patent_Citation
- Wipo

2. We have selected only the below information relevant for our analysis to determine the cited patents:
RELEVANT TABLES SELECTED

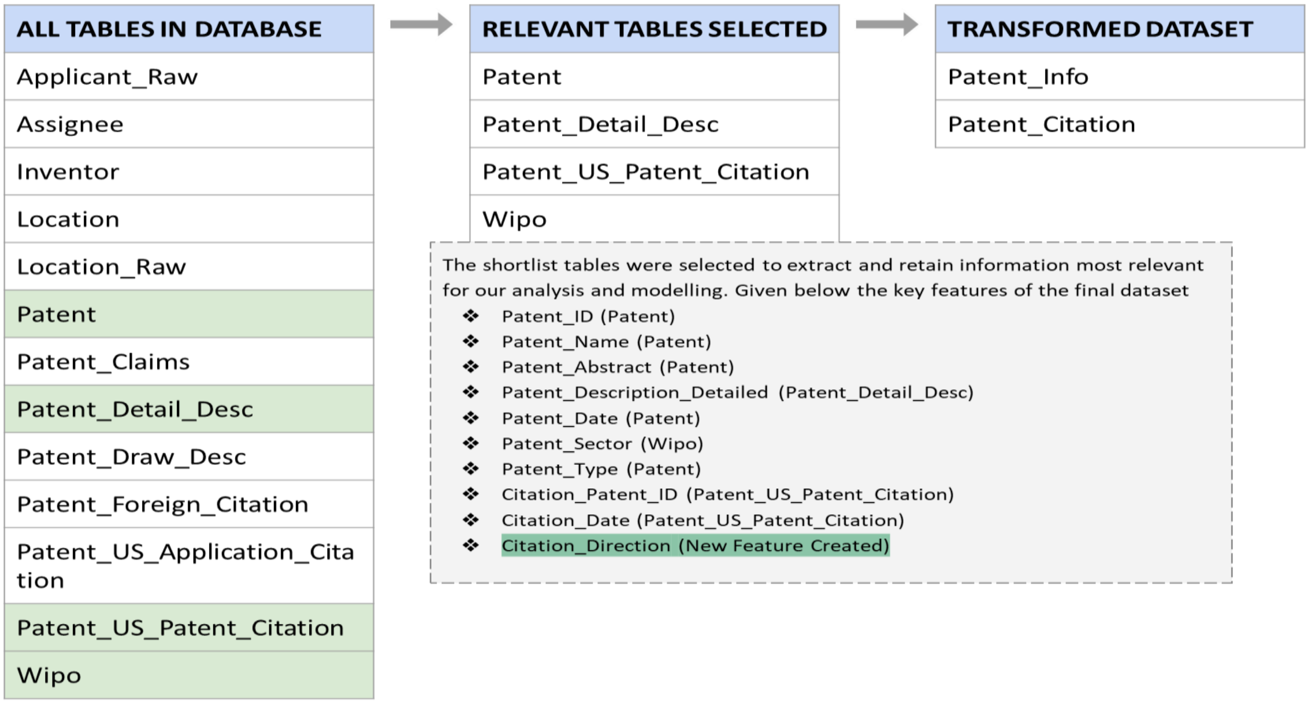
- Patent
- Patent_Detail_Desc
- Patent_US_Patent_Citation
- Wipo

3. The chosen tables were selected to extract and retain the most relevant information for our analysis and modeling. Here are the key features of the final dataset:

- Patent_ID (Patent)
- Patent_Name (Patent)

- Patent_Abstract (Patent)
- Patent_Description_Detailed (Patent_Detail_Desc)
- Patent_Date (Patent)
- Patent_Sector (Wipo)
- Patent_Type (Patent)
- Citation_Patent_ID (Patent_US_Patent_Citation)
- Citation_Date (Patent_US_Patent_Citation)

4. Our analysis will focus on a subset of the entire database consisting of 850,000 records, limited to a specific sector/industry.
5. The data has been directly sourced from our sponsors, providing proprietary access to comprehensive patent information.
6. We have also created a new feature to determine the direction of patent citation (backward or forward citation), crucial for our patent citation analysis, given our limited scope to backward citation.



ALL TABLES IN DATABASE
Applicant_Raw
Assignee
Inventor
Location
Location_Raw
Patent
Patent_Claims
Patent_Detail_Desc
Patent_Draw_Desc
Patent_Foreign_Citation
Patent_US_Application_Citation
Patent_US_Patent_Citation
Wipo

RELEVANT TABLES SELECTED
Patent
Patent_Detail_Desc
Patent_US_Patent_Citation
Wipo

TRANSFORMED DATASET
Patent_Info
Patent_Citation

The shortlist tables were selected to extract and retain information most relevant for our analysis and modelling. Given below the key features of the final dataset

- Patent_ID (Patent)
- Patent_Name (Patent)
- Patent_Abstract (Patent)
- Patent_Description_Detailed (Patent_Detail_Desc)
- Patent_Date (Patent)
- Patent_Sector (Wipo)
- Patent_Type (Patent)
- Citation_Patent_ID (Patent_US_Patent_Citation)
- Citation_Date (Patent_US_Patent_Citation)
- Citation_Direction (New Feature Created)

Fig: 6.1 Feature Selection

6.4 Text-Based Data Cleaning

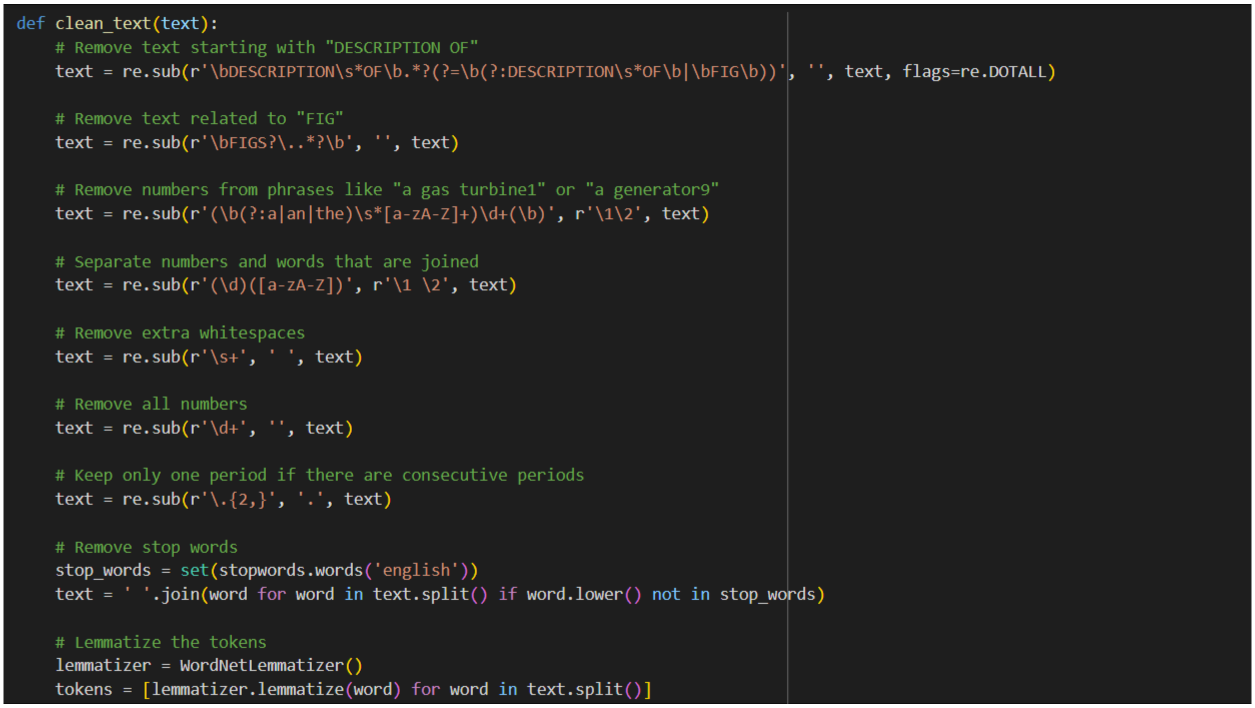
The following sections outline the steps involved in text-based data cleaning, ensuring the text data is refined and prepared for advanced analysis:

1. **Spell Checking**
   a. **Library Utilization:** The spellchecker library was employed to rectify spelling errors within the text.
   b. **Function Implementation:** A dedicated function, correct_spelling, was developed to perform spell checking and correction for individual words. This functionality was further extended to correct entire reviews via the correct_review function.

    c. **Batch Processing:** To efficiently handle large volumes of text, batch processing techniques were implemented.

2. **Text Preprocessing**
    a. **Spell Checking and Correction:** Functions were created to manage spell checking and correction for both individual words and complete reviews.
    b. **Batch Processing Enhancement:** The implementation of batch processing significantly improved the efficiency of text preprocessing operations.
3. **Specific Phrase Removal**
    a. **Identification and Removal:** The recurring phrase "detailed description of the invention" was identified, with an intention to remove or replace this specific phrase throughout the text.
4. **Number Removal**
    a. **Numerical Instances:** Specific numerical references such as "conditioner 120" were identified, with an intention to eliminate these instances from the text.
5. **Punctuation Removal**
    a. **Stripping Punctuation:** Punctuation marks including commas, periods, exclamation marks, and others were removed. This step is crucial to ensure punctuation does not interfere with subsequent text analysis tasks.
6. **Handling Contractions**
    a. **Expansion of Contractions:** Contractions such as "can't" were expanded to "cannot", and "don't" to "do not", enhancing consistency and clarity in the text data.
7. **Removing Special Characters**
    a. **Elimination of Non-Alphanumeric Characters:** Special characters, symbols, and other non-alphanumeric characters were removed from the text. These characters often do not contribute meaningfully to text analysis tasks and thus were excluded.

These steps collectively contribute to the cleaning and preprocessing of text data, making it more suitable for subsequent analysis tasks such as text summarization, sentiment analysis, and topic modeling.



```python
def clean_text(text):
    # Remove text starting with "DESCRIPTION OF"
    text = re.sub(r'\bDESCRIPTION\s*OF\b.*?(?=\b(?:DESCRIPTION\s*OF\b|\bFIG\b))', '', text, flags=re.DOTALL)

    # Remove text related to "FIG"
    text = re.sub(r'\bFIGS?\..*?\b', '', text)

    # Remove numbers from phrases like "a gas turbine1" or "a generator9"
    text = re.sub(r'(\b(?:a|an|the)\s*[a-zA-Z]+)\d+(\b)', r'\1\2', text)

    # Separate numbers and words that are joined
    text = re.sub(r'(\d)([a-zA-Z])', r'\1 \2', text)

    # Remove extra whitespaces
    text = re.sub(r'\s+', ' ', text)

    # Remove all numbers
    text = re.sub(r'\d+', '', text)

    # Keep only one period if there are consecutive periods
    text = re.sub(r'\.{2,}', '.', text)

    # Remove stop words
    stop_words = set(stopwords.words('english'))
    text = ' '.join(word for word in text.split() if word.lower() not in stop_words)

    # Lemmatize the tokens
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in text.split()]
```

Fig: 6.2 Text cleaning using RegEx

Chapter 7: Models Used

7.1 Why Chunking Was Used in Our Architecture?

Chunking was incorporated into our architecture to enhance the efficiency and accuracy of patent text analysis. This approach involves breaking down large text into smaller, more manageable segments while preserving the semantic integrity of the data. Given the complexity and length of patent documents, chunking helps in better handling and processing of text, ensuring that critical information is not lost during analysis.

7.1.1 Pros and Cons of Chunking

7.1.1.1 Pros

1. **Enhanced Efficiency:** Chunking allows for parallel processing, significantly speeding up the analysis of large texts. By dividing data into smaller chunks, computational resources can be utilized more effectively, leading to faster data processing times.
2. **Semantic Integrity:** Overlapping chunks help maintain the continuity of semantic information across segments. This overlap ensures that important context and meaning are preserved, which is crucial for accurate text analysis and recommendations.
3. **Scalability:** The chunking approach is highly scalable, allowing the processing system to handle growing datasets without a significant loss in performance. As the dataset size increases, chunking ensures that the system can continue to operate efficiently.

4. **Flexible Parameterization:** The chunk size and overlap can be adjusted based on the specific needs of the analysis, providing flexibility to optimize for different types of data and processing requirements.

7.1.1.2 Cons

1. **Potential Data Redundancy:** Overlapping chunks can introduce redundancy, as certain data points will appear in multiple chunks. While this helps preserve semantic integrity, it can also lead to increased storage requirements and potential duplication of processing efforts.
2. **Complexity in Implementation:** Implementing a chunking mechanism, especially with overlapping rows, adds complexity to the system. It requires careful planning and additional computational logic to ensure that chunks are created and processed correctly.
3. **Effectiveness Varies by Dataset:** The effectiveness of chunking can vary depending on the characteristics of the dataset and the specific processing tasks. For some datasets, chunking may not yield significant benefits and could even complicate the analysis process.

7.1.2 Evaluation of Chunking Approaches

Several chunking approaches were evaluated to determine the best method for our needs. These included **fixed-size chunking, dynamic chunking, and sliding window chunking**. Each method has its advantages and disadvantages, but due to the challenges of maintaining semantic coherence and avoiding information loss, we ultimately selected sliding window chunking.

7.1.2.1 Fixed-Size Chunking

Fixed-size chunking involves dividing the dataset into chunks of a predetermined size, with each chunk containing a fixed number of rows. This method is straightforward and easy to implement.

7.1.2.1.1 Why It Was Not Considered:

1. **Loss of Semantic Coherence:** Fixed-size chunking can disrupt the continuity of semantic information, especially in textual data where context is crucial. Important information might be split between chunks, leading to a loss of meaning and reducing the effectiveness of text analysis.
2. **Rigid Structure:** The inflexibility of fixed-size chunks does not account for the varying lengths of text and the natural breaks in data. This rigidity can result in either too much or too little information in each chunk, which complicates the analysis.
3. **Limited Scalability:** While simple to implement, fixed-size chunking does not scale well with datasets that have highly variable lengths of entries, making it less suitable for large and complex datasets like patent documents.

7.1.2.2 Dynamic Chunking

Dynamic chunking adjusts the size of each chunk based on the content, aiming to create chunks that align better with natural breaks in the data, such as sentences or paragraphs.

7.1.2.2.1 Why It Was Not Considered:

1. **Complex Implementation:** Dynamic chunking requires sophisticated algorithms to determine the optimal chunk boundaries based on content, which adds significant complexity to the implementation process.

2. **Inconsistent Chunk Sizes:** This method results in chunks of varying sizes, which can complicate the processing and analysis stages. Managing and analyzing chunks of different lengths can be more challenging and less efficient than working with uniformly sized chunks.
3. **Potential for Semantic Loss:** Despite its aim to preserve semantic integrity, dynamic chunking can still fail to capture the full context if the algorithm does not accurately identify the optimal break points, leading to potential gaps in the data.

7.1.2.3 Sliding Window Chunking:

Sliding window chunking involves creating overlapping chunks of data by sliding a window of a fixed size over the dataset. Each window overlaps with the previous one by a specified number of rows, ensuring that critical context is preserved across chunks.

7.1.2.3.1 Why Sliding Window Chunking Was Chosen

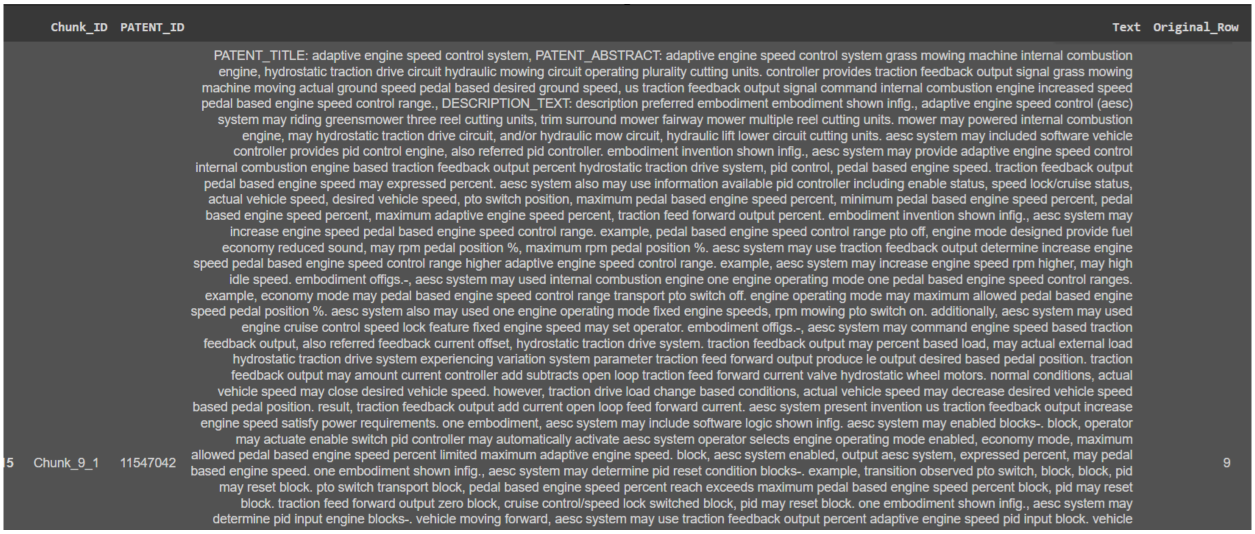
1. **Enhanced Semantic Integrity:** By overlapping chunks, sliding window chunking ensures that important context is maintained across boundaries. This method minimizes the risk of losing semantic coherence, which is essential for accurate text analysis.
2. **Balanced Flexibility:** Sliding window chunking provides a balance between fixed-size and dynamic chunking. It maintains uniform chunk sizes for efficient processing while allowing overlap to capture semantic continuity.
3. **Improved Accuracy:** The overlap between chunks helps preserve the meaning of the text, ensuring that key information is not fragmented. This results in more accurate analysis and better outcomes for tasks such as summarization and recommendation.
4. **Scalability and Efficiency:** Sliding window chunking is scalable and can be adjusted easily by changing the window size and overlap. This flexibility allows it to handle large datasets efficiently, making it suitable for complex datasets like patent documents.

7.1.3 Trade-Off

The primary trade-off with sliding window chunking is the introduction of redundancy. Some data points are duplicated in multiple chunks to preserve semantics, which can lead to increased storage and processing requirements. However, this redundancy is a necessary compromise to ensure that the context and meaning of the data are accurately maintained throughout the analysis.

7.1.4 Conclusion

Chunking, and specifically sliding window chunking, was adopted to enhance the efficiency and effectiveness of our patent analysis system. Despite the trade-offs, this approach provides a robust solution to manage and process large datasets while maintaining the integrity of the semantic information critical to accurate analysis and recommendations.



	Chunk_ID	PATENT_ID	Text	Original_Row
5	Chunk_9_1	11547042	PATENT_TITLE: adaptive engine speed control system, PATENT_ABSTRACT: adaptive engine speed control system grass mowing machine internal combustion engine, hydrostatic traction drive circuit hydraulic mowing circuit operating plurality cutting units. controller provides traction feedback output signal grass mowing machine moving actual ground speed pedal based desired ground speed, us traction feedback output signal command internal combustion engine increased speed pedal based engine speed control range., DESCRIPTION_TEXT: description preferred embodiment embodiment shown infig., adaptive engine speed control (aesc) system may riding greensmower three reel cutting units, trim surround mower fairway mower multiple reel cutting units. mower may powered internal combustion engine, may hydrostatic traction drive circuit, and/or hydraulic mow circuit, hydraulic lift lower circuit cutting units. aesc system may included software vehicle controller provides pid control engine, also referred pid controller. embodiment invention shown infig., aesc system may provide adaptive engine speed control internal combustion engine based traction feedback output percent hydrostatic traction drive system, pid control, pedal based engine speed. traction feedback output pedal based engine speed may expressed percent. aesc system also may use information available pid controller including enable status, speed lock/cruise status, actual vehicle speed, desired vehicle speed, pto switch position, maximum pedal based engine speed percent, minimum pedal based engine speed percent, pedal based engine speed percent, maximum adaptive engine speed percent, traction feed forward output percent. embodiment invention shown infig., aesc system may increase engine speed pedal based engine speed control range. example, pedal based engine speed control range pto off, engine mode designed provide fuel economy reduced sound, may rpm pedal position %, maximum rpm pedal position %. aesc system may use traction feedback output determine increase engine speed pedal based engine speed control range higher adaptive engine speed control range. example, aesc system may increase engine speed rpm higher, may high idle speed. embodiment offigs.-, aesc system may used internal combustion engine one engine operating mode one pedal based engine speed control ranges. example, economy mode may pedal based engine speed control range transport pto switch off. engine operating mode may maximum allowed pedal based engine speed pedal position %. aesc system also may used one engine operating mode fixed engine speeds, rpm mowing pto switch on. additionally, aesc system may used engine cruise control speed lock feature fixed engine speed may set operator. embodiment offigs.-, aesc system may command engine speed based traction feedback output, also referred feedback current offset, hydrostatic traction drive system. traction feedback output may percent based load, may actual external load hydrostatic traction drive system experiencing variation system parameter traction feed forward output produce le output desired based pedal position. traction feedback output may amount current controller add subtracts open loop traction feed forward current valve hydrostatic wheel motors. normal conditions, actual vehicle speed may close desired vehicle speed. however, traction drive load change based conditions, actual vehicle speed may decrease desired vehicle speed based pedal position. result, traction feedback output add current open loop feed forward current. aesc system present invention us traction feedback output increase engine speed satisfy power requirements. one embodiment, aesc system may include software logic shown infig. aesc system may enabled blocks-. block, operator may actuate enable switch pid controller may automatically activate aesc system operator selects engine operating mode enabled, economy mode, maximum allowed pedal based engine speed percent limited maximum adaptive engine speed. block, aesc system enabled, output aesc system, expressed percent, may pedal based engine speed. one embodiment shown infig., aesc system may determine pid reset condition blocks-. example, transition observed pto switch, block, block, pid may reset block. pto switch transport block, pedal based engine speed percent reach exceeds maximum pedal based engine speed percent block, pid may reset block. traction feed forward output zero block, cruise control/speed lock switched block, pid may reset block. one embodiment shown infig., aesc system may determine pid input engine blocks-. vehicle moving forward, aesc system may use traction feedback output percent adaptive engine speed pid input block. vehicle	9

Fig: 7.1 Chunk creation

7.2 Utilizing RAG (Retrieval-Augmented Generation) for Patent Summarization: Evaluation and Selection

In the realm of patent analysis and summarization, the adoption of advanced techniques like Retrieval-Augmented Generation (RAG) has become increasingly prevalent. RAG combines the strengths of retrieval-based methods with generative models to produce concise and relevant summaries of patent documents. In our endeavor to enhance cited patent analysis and recommendations, we evaluated several RAG models and ultimately selected Google Pegasus for its superior performance and adaptability to our specific domain needs.

7.2.1 Why RAG was Used?

RAG was chosen as our approach for patent summarization due to several key advantages it offers:

1. **Integration of Retrieval and Generation:** RAG seamlessly integrates retrieval and generation mechanisms, allowing for the extraction of relevant information from patent documents while also generating concise summaries. This dual approach ensures that the produced summaries are grounded in the context of the original text, enhancing their relevance and accuracy.
2. **Flexibility and Customization:** RAG models can be fine-tuned and customized to suit specific domain requirements and data characteristics. By training RAG models on domain-specific datasets, such as patent data, we can optimize performance and ensure that the generated summaries meet the unique needs of patent researchers and analysts.
3. **Semantic Understanding:** RAG models, including Google Pegasus, leverage advanced language understanding capabilities to capture the semantic meaning and context of patent documents. This enables more nuanced summarization that goes beyond simple keyword extraction, resulting in summaries that better reflect the essence of the original text.

7.2.2 Pros and Cons of RAG:

7.2.2.1 Pros

1. **Contextual Relevance:** RAG models generate summaries that maintain the context and coherence of the original text, enhancing the relevance and usability of the generated summaries.
2. **Customizability:** RAG models can be fine-tuned and adapted to specific domains and datasets, allowing for improved performance and relevance in summarization tasks.
3. **Semantic Understanding:** RAG models excel at capturing the semantic meaning and nuances of text, enabling more accurate and nuanced summarization compared to traditional keyword-based approaches.

7.2.2.2 Cons

1. **Computational Complexity:** Training and fine-tuning RAG models can be computationally intensive, requiring significant computational resources and expertise.
2. **Data Dependency:** The performance of RAG models is highly dependent on the quality and diversity of the training data, necessitating large and diverse datasets for optimal performance.
3. **Interpretability:** The inner workings of RAG models can be complex and difficult to interpret, making it challenging to understand how the model arrives at its summarization decisions.

7.2.3 Selection of Google Pegasus

In the quest to enhance patent search and summarization, we evaluated multiple models to identify the most suitable solution. After careful consideration, we selected Google Pegasus for its superior performance and alignment with our specific requirements. Here's a detailed explanation of why Pegasus was chosen over other models and the benefits it provided post fine-tuning.

7.2.3.1 Why We Didn't Use Other Models?

1. **Blaise-g/longt5_tglobal_large_scitldr:**
   a) **Limitations:** Though effective for scientific document summarization, it struggled with the complexity and specificity of patent documents.
   b) **Performance:** Lower relevance in generated summaries compared to Pegasus.

2. **0x70DA/t5-v1_1-base-finetuned-sci_summ:**
   a) **Limitations:** This model showed limitations in handling the intricate details and technical jargon of patents.
   b) **Performance:** Precision and coherence of summaries were not on par with Pegasus.

3. **sambydlo/bart-large-scientific-lay-summarisation:**
   a) **Limitations:** Primarily designed for lay summarization, which didn't translate well to the detailed nature of patent documents.
   b) **Performance:** Generated summaries were overly simplified and lacked necessary technical depth.

4. **sambydlo/scientific_abstract_simplification-scientific-lay-summarise:**

   a) **Limitations:** Similar to the above, this model was too simplified for the patent domain.
   b) **Performance:** Summaries were not sufficiently detailed or relevant.

5. **pszemraj/long-t5-tglobal-base-sci-simplify:**
   a) **Limitations:** Designed more for scientific document simplification rather than technical patent summarization.
   b) **Performance:** Lacked precision in maintaining patient-specific information.

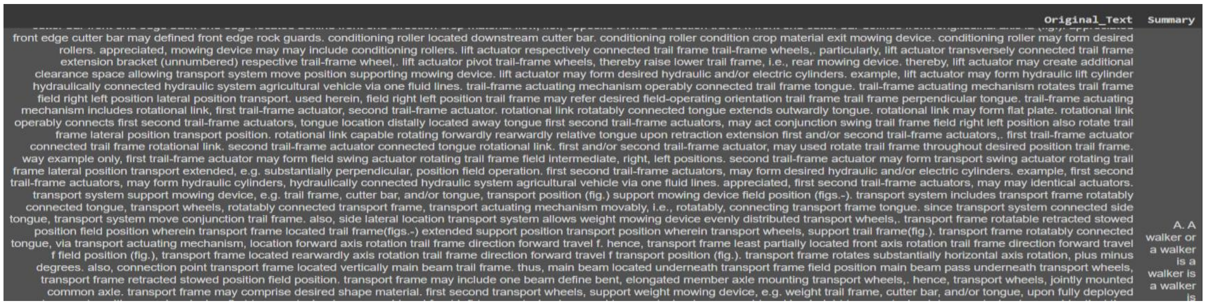
6. **pszemraj/long-t5-tglobal-base-sci-simplify-elife:**
   a) **Limitations:** While it improved readability, it did so at the cost of omitting critical technical details.
   b) **Performance:** Not suitable for retaining the necessary depth required for patent documents.

7. **chizhikchi/sci-five-radsum23:**
   a) **Limitations:** Focused on radiology summaries, this model did not generalize well to the broader patent domain.
   b) **Performance:** Context and relevance in patent summarization were suboptimal.

8. **pszemraj/long-t5-tglobal-xl-sci-simplify-elife:**
   a) **Limitations:** Although capable, the model's focus on scientific simplification was not aligned with the detailed nature of patents.
   b) **Performance:** Generated summaries were lacking in patent-specific details.

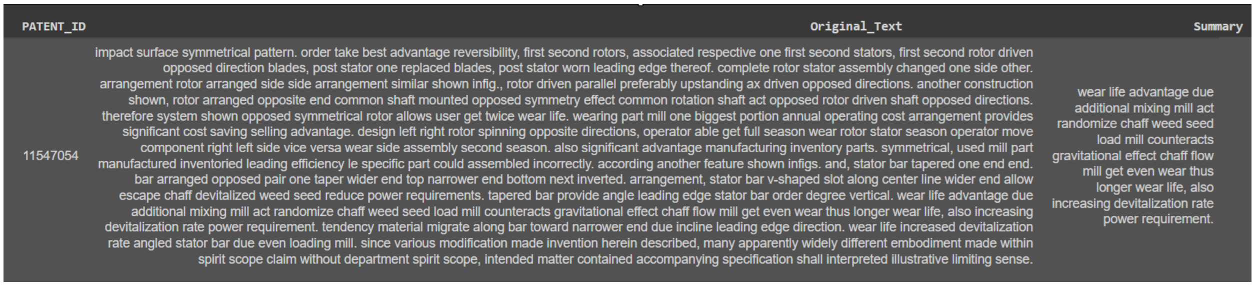
9. **AhiyaB/mt5-small-finetuned-Big-Patent-h:**
   a) **Limitations:** Performance was inconsistent, and summaries often missed key elements of the patents.
   b) **Performance:** Overall quality and coherence were lower than those produced by Pegasus.

10. **anferico/bert-for-patents:**
   a) **Limitations:** Designed for classification rather than summarization, making it less suitable for our needs.
   b) **Performance:** Did not perform well in generating coherent and concise summaries.



Original_Text	Summary
front edge cutter bar may defined front edge rock guards. conditioning roller located downstream cutter bar. conditioning roller condition crop material exit mowing device. conditioning roller may form desired rollers. appreciated, mowing device may may include conditioning rollers. lift actuator respectively connected trail frame trail-frame wheels,. particularly, lift actuator transversely connected trail frame extension bracket (unnumbered) respective trail-frame wheel,. lift actuator pivot trail-frame wheels, thereby raise lower trail frame, i.e., rear mowing device. thereby, lift actuator may create additional clearance space allowing transport system move position supporting mowing device. lift actuator may form desired hydraulic and/or electric cylinders. example, lift actuator may form hydraulic lift cylinder hydraulically connected hydraulic system agricultural vehicle via one fluid lines. trail-frame actuating mechanism operably connected trail frame tongue. trail-frame actuating mechanism rotates trail frame field right left position lateral position transport. used herein, field right left position trail frame may refer desired field-operating orientation trail frame trail frame perpendicular tongue. trail-frame actuating mechanism includes rotational link, first trail-frame actuator, second trail-frame actuator. rotational link rotatably connected tongue extends outwardly tongue. rotational link may form flat plate. rotational link operably connects first second trail-frame actuators, tongue location distally located away tongue first second trail-frame actuators, may act conjunction swing trail frame field right left position also rotate trail frame lateral position transport position. rotational link capable rotating forwardly rearwardly relative tongue upon retraction extension first and/or second trail-frame actuators,. first trail-frame actuator connected trail frame rotational link. second trail-frame actuator connected tongue rotational link. first and/or second trail-frame actuator, may used rotate trail frame throughout desired position trail frame. way example only, first trail-frame actuator may form field swing actuator rotating trail frame field intermediate, right, left positions. second trail-frame actuator may form transport swing actuator rotating trail frame lateral position transport extended, e.g. substantially perpendicular, position field operation. first second trail-frame actuators, may form desired hydraulic and/or electric cylinders. example, first second trail-frame actuators, may form hydraulic cylinders, hydraulically connected hydraulic system agricultural vehicle via one fluid lines. appreciated, first second trail-frame actuators, may may identical actuators. transport system support mowing device, e.g. trail frame, cutter bar, and/or tongue, transport position (fig.) support mowing device field position (figs.-). transport system includes transport frame rotatably connected tongue, transport wheels, rotatably connected transport frame, transport actuating mechanism movably, i.e., rotatably, connecting transport frame tongue. since transport system connected side tongue, transport system move conjunction trail frame. also, side lateral location transport system allows weight mowing device evenly distributed transport wheels,. transport frame rotatable retracted stowed position field position wherein transport frame located trail frame(figs.-) extended support position transport position wherein transport wheels, support trail frame(fig.). transport frame rotatably connected tongue, via transport actuating mechanism, location forward axis rotation trail frame direction forward travel f. hence, transport frame least partially located front axis rotation trail frame direction forward travel f field position (fig.), transport frame located rearwardly axis rotation trail frame direction forward travel f transport position (fig.). transport frame rotates substantially horizontal axis rotation, plus minus degrees. also, connection point transport frame located vertically main beam trail frame. thus, main beam located underneath transport frame field position main beam pass underneath transport wheels, transport frame retracted stowed position field position. transport frame may include one beam define bent, elongated member axle mounting transport wheels,. hence, transport wheels, jointly mounted common axle. transport frame may comprise desired shape material. first second transport wheels, support weight mowing device, e.g. weight trail frame, cutter bar, and/or tongue, upon fully deployed	A. A walker or a walker is a walker is a walker is

Fig: 7.2 Text summarization with other RAG model (RAG tokenizer)



PATENT_ID	Original_Text	Summary
11547054	impact surface symmetrical pattern. order take best advantage reversibility, first second rotors, associated respective one first second stators, first second rotor driven opposed direction blades, post stator one replaced blades, post stator worn leading edge thereof. complete rotor stator assembly changed one side other. arrangement rotor arranged side side arrangement similar shown infig., rotor driven parallel preferably upstanding ax driven opposed directions. another construction shown, rotor arranged opposite end common shaft mounted opposed symmetry effect common rotation shaft act opposed rotor driven shaft opposed directions. therefore system shown opposed symmetrical rotor allows user get twice wear life. wearing part mill one biggest portion annual operating cost arrangement provides significant cost saving selling advantage. design left right rotor spinning opposite directions, operator able get full season wear rotor stator season operator move component right left side vice versa wear side assembly second season. also significant advantage manufacturing inventory parts. symmetrical, used mill part manufactured inventoried leading efficiency le specific part could assembled incorrectly. according another feature shown infigs. and, stator bar tapered one end end. bar arranged opposed pair one taper wider end top narrower end bottom next inverted. arrangement, stator bar v-shaped slot along center line wider end allow escape chaff devitalized weed seed reduce power requirements. tapered bar provide angle leading edge stator bar order degree vertical. wear life advantage due additional mixing mill act randomize chaff weed seed load mill counteracts gravitational effect chaff flow mill get even wear thus longer wear life, also increasing devitalization rate power requirement. tendency material migrate along bar toward narrower end due incline leading edge direction. wear life increased devitalization rate angled stator bar due even loading mill. since various modification made invention herein described, many apparently widely different embodiment made within spirit scope claim without department spirit scope, intended matter contained accompanying specification shall interpreted illustrative limiting sense.	wear life advantage due additional mixing mill act randomize chaff weed seed load mill counteracts gravitational effect chaff flow mill get even wear thus longer wear life, also increasing devitalization rate power requirement.

Fig: 7.3 Text summarization with Pegasus RAG mode

Fine tuned pegasus on sample set

7.2.3.2 Why Pegasus Was Used

- Pegasus is a transformer encoder-decoder model designed specifically for text generation tasks, particularly abstractive summarization.
- It utilizes a novel pre-training objective, Gap Sentences Generation (GSG), where entire sentences are masked, and the model is trained to generate these sentences based on the context provided by the rest of the text.
- This pre-training makes Pegasus particularly adept at generating coherent and concise summaries, capturing the essence of the input text effectively.

7.2.3.3 Key Advantages of Pegasus

1. **Pre-training on Patent Data:** Pegasus has been pre-trained on a large corpus of patent data, making it highly proficient in understanding and summarizing patent documents.
2. **GSG Objective:** The GSG objective helps the model learn to compress information and generate coherent summaries, making it ideal for summarization tasks.
3. **Fine-tuning Capability:** Fine-tuning Pegasus involves adjusting a model already optimized for summarization, enhancing its efficiency and effectiveness for specific tasks.

7.2.3.4 Fine-tuning Pegasus

To further improve the performance of Pegasus in our domain, we fine-tuned it on our specific patent dataset. This process involved:

1. **Hyper-parameter Tuning:** Adjusting parameters to optimize model performance on our specific dataset.

2. **Evaluation Metrics:** Calculating metrics such as ROUGE, BERT, and BLEU scores to assess semantic loss and contextual differences between original and generated summaries.

7.2.3.5 Added Benefits of Fine-tuning Pegasus

1. **Enhanced Relevance and Accuracy:** Fine-tuning improved the relevance and accuracy of the summaries, ensuring they captured the essential information from the patents.
2. **Better Semantic Understanding:** The model's understanding of patent-specific terminology and context was significantly enhanced.
3. **Improved Evaluation Scores:** Post fine-tuning, Pegasus showed improved ROUGE, BERT, and BLEU scores, indicating higher quality and more faithful summaries.

Overall, **Google Pegasus emerged as the optimal choice for our RAG model,** providing a powerful combination of performance, adaptability, and domain relevance. Fine-tuning further amplified its capabilities, making it an invaluable tool for patent summarization and analysis.

7.3 Why Pegasus was considered over BERT: Embeddings generated through Pegasus

1. **Architecture Differences:**
   BERT
   1.1 BERT is a bidirectional transformer encoder designed primarily for natural language understanding tasks such as classification, named entity recognition, and question answering.
   1.2 It generates context-aware embeddings by considering the entire sentence or document, but it doesn't generate new sequences (e.g., summaries or translations).
   1.3 BERT embeddings are typically used as input to downstream tasks where the embeddings represent the context of each token in the input.
   Pegasus
   1.1 Pegasus is a transformer encoder-decoder model designed for text generation tasks, particularly abstractive summarization.
   1.2 It uses a novel pre-training objective tailored for summarization, where entire sentences are masked and the model is trained to generate these sentences based on the context provided by the rest of the text.
   1.3 Pegasus can generate new sequences from input text, making it more suitable for tasks like summarization, paraphrasing, and other generation tasks.
2. **Pre-training Objectives:**
   BERT:
   2.1 BERT uses the Masked Language Model (MLM) and Next Sentence Prediction (NSP) objectives.
   2.2 MLM involves masking 15% of the input tokens and training the model to predict these masked tokens based on their context.
   2.3 NSP involves predicting whether a given sentence follows another in the text, which helps the model understand sentence relationships.
   Pegasus:
   2.1 Pegasus uses the Gap Sentences Generation (GSG) pre-training objective.

2.2 GSG involves masking entire sentences and training the model to generate these sentences, simulating the summarization task more directly.

2.3 This objective helps the model learn to compress information and generate coherent summaries, making it highly effective for summarization and related tasks.

3. **Fine-tuning for Summarization:**

   BERT:

   3.1 While BERT can be fine-tuned for summarization tasks, it requires additional mechanisms (e.g., adding a sequence-to-sequence layer) since its original design is not for generation.

   3.2 Fine-tuning BERT for generation tasks is less straightforward and often less effective compared to models designed for generation.

   Pegasus:

   3.1 Pegasus, being designed for summarization, directly supports fine-tuning for text generation tasks.

   3.2 Fine-tuning Pegasus involves adjusting a model that is already optimized for tasks like summarization, making it more efficient and effective.

4. **Embedding Use Cases:**

   BERT:

   4.1 BERT embeddings are excellent for understanding the context within sentences and performing tasks like classification and sequence labeling.

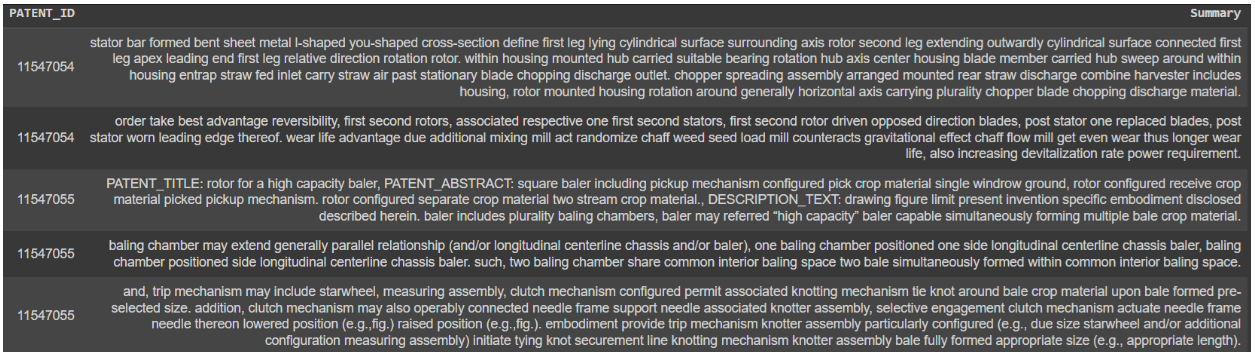
   4.2 They capture rich contextual information for individual tokens but do not inherently provide a mechanism for generating new text.

   Pegasus:

   4.1 Pegasus embeddings are optimized for generating coherent and contextually appropriate new text sequences.

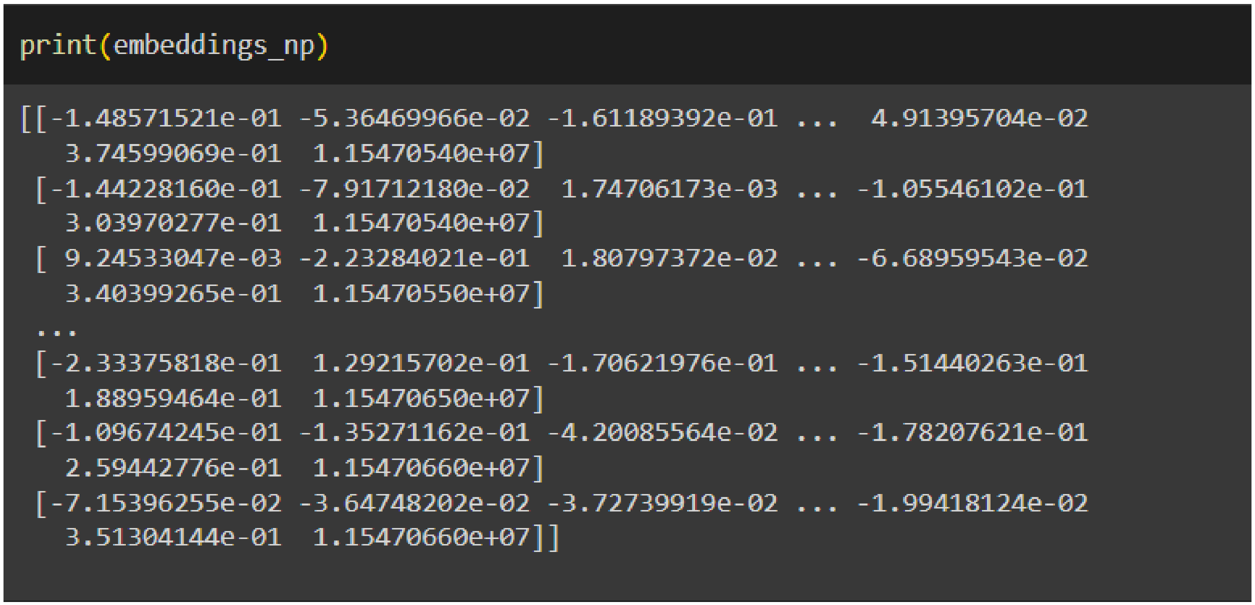
   4.2 For tasks that involve generating summaries or new text, Pegasus embeddings are more suitable because they are trained to understand and generate text sequences.

Conclusion: Pegasus's architecture and pre-training objectives provide a significant advantage to create embeddings that are used for generating new text sequences (e.g., summaries).



PATENT_ID	Summary
11547054	stator bar formed bent sheet metal l-shaped you-shaped cross-section define first leg lying cylindrical surface surrounding axis rotor second leg extending outwardly cylindrical surface connected first leg apex leading end first leg relative direction rotation rotor. within housing mounted hub carried suitable bearing rotation hub axis center housing blade member carried hub sweep around within housing entrap straw fed inlet carry straw air past stationary blade chopping discharge outlet. chopper spreading assembly arranged mounted rear straw discharge combine harvester includes housing, rotor mounted housing rotation around generally horizontal axis carrying plurality chopper blade chopping discharge material.
11547054	order take best advantage reversibility, first second rotors, associated respective one first second stators, first second rotor driven opposed direction blades, post stator one replaced blades, post stator worn leading edge thereof. wear life advantage due additional mixing mill act randomize chaff weed seed load mill counteracts gravitational effect chaff flow mill get even wear thus longer wear life, also increasing devitalization rate power requirement.
11547055	PATENT_TITLE: rotor for a high capacity baler, PATENT_ABSTRACT: square baler including pickup mechanism configured pick crop material single windrow ground, rotor configured receive crop material picked pickup mechanism. rotor configured separate crop material two stream crop material., DESCRIPTION_TEXT: drawing figure limit present invention specific embodiment disclosed described herein. baler includes plurality baling chambers, baler may referred “high capacity” baler capable simultaneously forming multiple bale crop material.
11547055	baling chamber may extend generally parallel relationship (and/or longitudinal centerline chassis and/or baler), one baling chamber positioned one side longitudinal centerline chassis baler, baling chamber positioned side longitudinal centerline chassis baler. such, two baling chamber share common interior baling space two bale simultaneously formed within common interior baling space.
11547055	and, trip mechanism may include starwheel, measuring assembly, clutch mechanism configured permit associated knotting mechanism tie knot around bale crop material upon bale formed pre-selected size. addition, clutch mechanism may also operably connected needle frame support needle associated knotter assembly, selective engagement clutch mechanism actuate needle frame needle thereon lowered position (e.g.,fig.) raised position (e.g.,fig.). embodiment provide trip mechanism knotter assembly particularly configured (e.g., due size starwheel and/or additional configuration measuring assembly) initiate tying knot securement line knotting mechanism knotter assembly bale fully formed appropriate size (e.g., appropriate length).

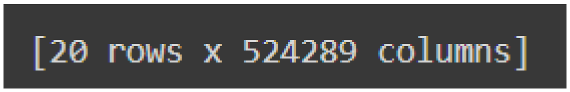
Fig: 7.4 Text before embeddings



```
print(embeddings_np)
```

```
[[-1.48571521e-01 -5.36469966e-02 -1.61189392e-01 ...  4.91395704e-02
   3.74599069e-01  1.15470540e+07]
 [-1.44228160e-01 -7.91712180e-02  1.74706173e-03 ... -1.05546102e-01
   3.03970277e-01  1.15470540e+07]
 [ 9.24533047e-03 -2.23284021e-01  1.80797372e-02 ... -6.68959543e-02
   3.40399265e-01  1.15470550e+07]
 ...
 [-2.33375818e-01  1.29215702e-01 -1.70621976e-01 ... -1.51440263e-01
   1.88959464e-01  1.15470650e+07]
 [-1.09674245e-01 -1.35271162e-01 -4.20085564e-02 ... -1.78207621e-01
   2.59442776e-01  1.15470660e+07]
 [-7.15396255e-02 -3.64748202e-02 -3.72739919e-02 ... -1.99418124e-02
   3.51304144e-01  1.15470660e+07]]
```

Fig: 7.5 Text converted to embeddings



```
[20 rows x 524289 columns]
```

Fig: 7.6 Number of embeddings data dimension

7.4 FAISS

Utilizing FAISS Database for Patent Summarization: Evaluation and Selection

In the continued effort to enhance our patent summarization process, we evaluated various database solutions to efficiently manage and retrieve relevant information. After thorough analysis, we chose the FAISS (Facebook AI Similarity Search) database over Chroma for its superior performance in handling our specific needs.

7.4.1 Why FAISS was Used?

FAISS was selected for our patent summarization system due to several key advantages it offers:

High Efficiency and Scalability: FAISS is optimized for large-scale similarity search, enabling efficient indexing and retrieval of large patent datasets. Its scalability ensures that our growing database of patents can be managed effectively without compromising on retrieval speed or accuracy.

Advanced Similarity Search: FAISS utilizes state-of-the-art algorithms for nearest neighbor search, allowing for precise retrieval of relevant patent documents. This capability is crucial for generating accurate and contextually relevant summaries.

Customizability and Flexibility: FAISS provides extensive customization options, enabling us to fine-tune the indexing and search parameters to best suit our domain-specific requirements. This flexibility allows for enhanced performance tailored to our patent data characteristics.

7.4.2 Pros and Cons of FAISS

7.4.2.1 Pros

- **Efficient Retrieval: FAISS is designed to handle high-dimensional data and perform fast similarity searches, which is essential for managing extensive patent datasets.**
- **Scalability:** The ability of FAISS to scale with the dataset size ensures that our system remains efficient and effective as the number of patents grows.
- **Customizable Indexes:** FAISS supports various indexing methods, including flat (brute-force) and more advanced approaches like Inverted File Indexes (IVF) and HNSW (Hierarchical Navigable Small World). This allows for optimized performance based on specific needs.

7.4.2.2 Cons

- **Complexity:** Implementing and optimizing FAISS requires a deep understanding of its indexing and search mechanisms, necessitating expertise and time for setup and maintenance.
- **Memory Usage:** Depending on the indexing method used, FAISS can be memory-intensive, which might require significant computational resources for optimal performance.
- **Hardware Requirements:** To fully leverage FAISS's capabilities, robust hardware infrastructure may be necessary, particularly for handling large-scale datasets and complex indexing.

7.4.3 Selection of FAISS over Chroma

In our evaluation, we compared FAISS with Chroma, a potential alternative for our patent summarization needs. Here’s a detailed explanation of why FAISS was chosen over Chroma and the benefits it provided post-implementation.

7.4.3.1 Why We Didn't Use Chroma?

- **Performance Limitations:** Chroma, while effective for certain applications, did not match FAISS's performance in large-scale similarity searches and efficient retrieval of high-dimensional data.
- **Scalability Issues:** Chroma's scalability was found to be less robust compared to FAISS, making it less suitable for our growing patent database.
- **Customization Constraints:** Chroma offered fewer customization options for indexing and retrieval, limiting our ability to optimize the system for our specific domain requirements.

7.4.3.2 Why FAISS Was Used?

FAISS is a comprehensive library developed by Facebook AI Research that excels in high-dimensional similarity searches. Here are the key reasons for its selection:

Advanced Indexing Techniques: FAISS supports multiple indexing methods, allowing us to choose the most suitable one for our data characteristics. This flexibility ensures optimal performance in terms of speed and accuracy.

Efficient Use of Resources: Despite its complexity, FAISS is highly efficient in utilizing computational resources, enabling us to handle large datasets without excessive hardware investments.

Strong Community and Support: FAISS has a robust community and extensive documentation, providing valuable support and resources for implementation and optimization.

7.4.4 Implementing FAISS

To fully leverage FAISS for our patent summarization needs, we undertook the following steps:

Index Selection and Configuration: Based on our data characteristics, we selected the most appropriate indexing method (e.g., IVF, HNSW) and fine-tuned the parameters for optimal performance.

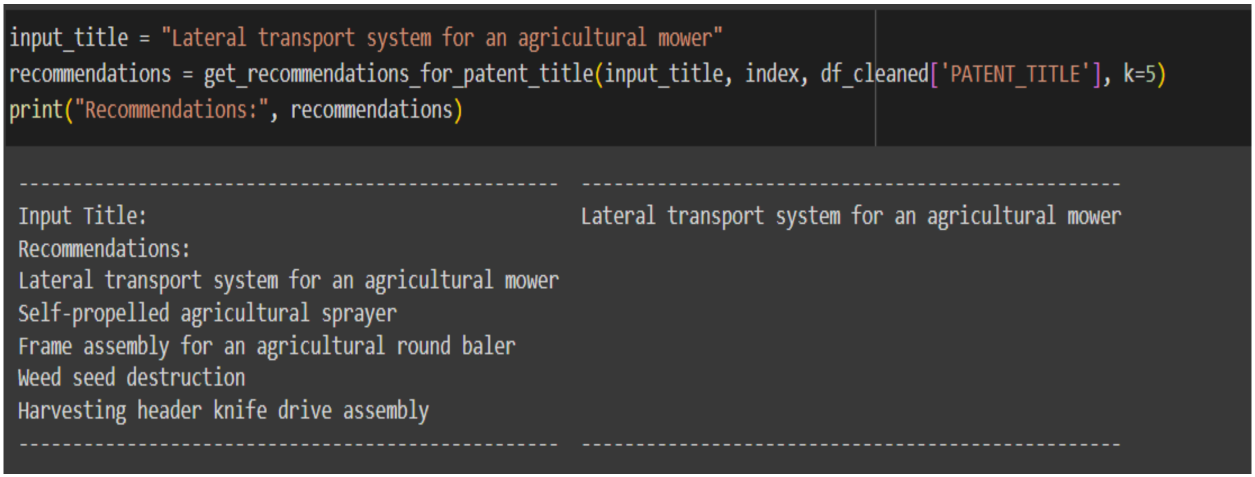
Integration with Summarization Pipeline: FAISS was seamlessly integrated into our summarization pipeline, ensuring efficient retrieval of relevant patents for summarization tasks.

Performance Monitoring and Optimization: Continuous monitoring and optimization were carried out to ensure that FAISS met our performance expectations, adjusting parameters as necessary to handle the growing dataset.

7.4.5 Benefits of Using FAISS

- **Enhanced Retrieval Speed and Accuracy:** FAISS significantly improved the speed and accuracy of retrieving relevant patent documents, directly contributing to more precise and relevant summaries.
- **Scalable Solution:** The scalability of FAISS ensured that our system remained efficient and effective as our patent database expanded.
- **Optimized Resource Utilization:** Through careful configuration and optimization, FAISS provided efficient use of computational resources, balancing performance with hardware constraints.

Overall, FAISS emerged as the optimal database solution for our patent summarization needs, offering a powerful combination of performance, scalability, and flexibility. Its implementation has greatly enhanced our ability to generate accurate and relevant patent summaries, supporting our overarching goal of improving cited patent analysis and recommendations.



```
input_title = "Lateral transport system for an agricultural mower"
recommendations = get_recommendations_for_patent_title(input_title, index, df_cleaned['PATENT_TITLE'], k=5)
print("Recommendations:", recommendations)
```

```
-------------------------------------------------  -------------------------------------------------
Input Title:                                       Lateral transport system for an agricultural mower
Recommendations:
Lateral transport system for an agricultural mower
Self-propelled agricultural sprayer
Frame assembly for an agricultural round baler
Weed seed destruction
Harvesting header knife drive assembly
-------------------------------------------------  -------------------------------------------------
```

Fig: 7.6 Final recommendation list

Chapter 8. Model Deployment and Evaluation

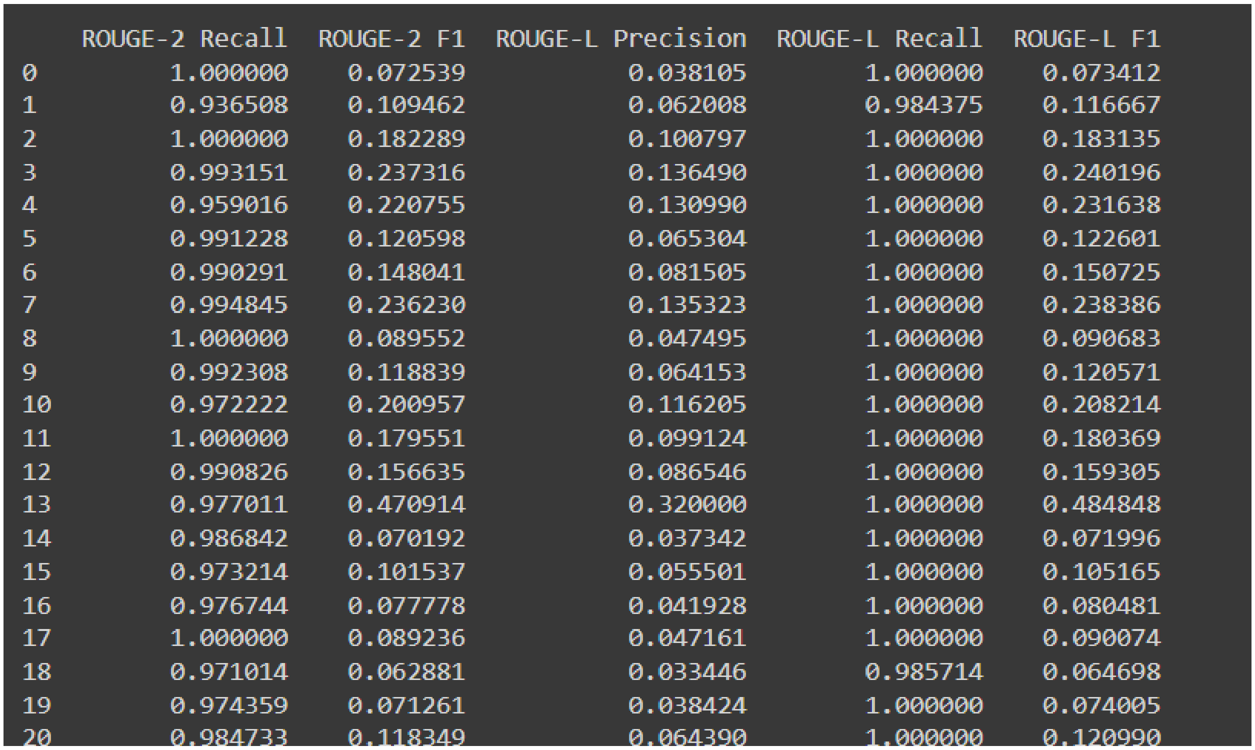
In this section, we consider the used and applied models for automated patent citation analysis and their measures. The main task is to apply large language models for generating embedding features and searching the related similarity.
The model was deployed on an EC2 instance with a Python script developed to process user inputs and provide patent recommendations as potential citations. EC2 was chosen for its scalability, reliability, and cost-effectiveness. This cloud-based service allows for easy adjustments to computing power based on demand, ensuring efficient and straightforward deployment and operation without adding unnecessary complexity. Additionally, EC2's robust security features and extensive support make it an ideal choice for deploying and maintaining machine learning models.

Table 8.1- Model Description

Model	Brief approach	Challenges
Google PEGASUS	Encode the semantic meaning into vector representations of the textual information.	Computationally intense involves a lot of memory use and central processing units.
FAISS (Facebook AI Similarity Search)	Utilizes vector embeddings for faster similarity searches establishing indices for effective data storage and retrieval.	Managing big datasets; finding the balance between accuracy and speed when optimizing index parameters.

These models were selected since they could accurately manage semantic context along with intelligently matching the results. This is very essential to support the patent citation analysis which is the prime focus of this simulation.

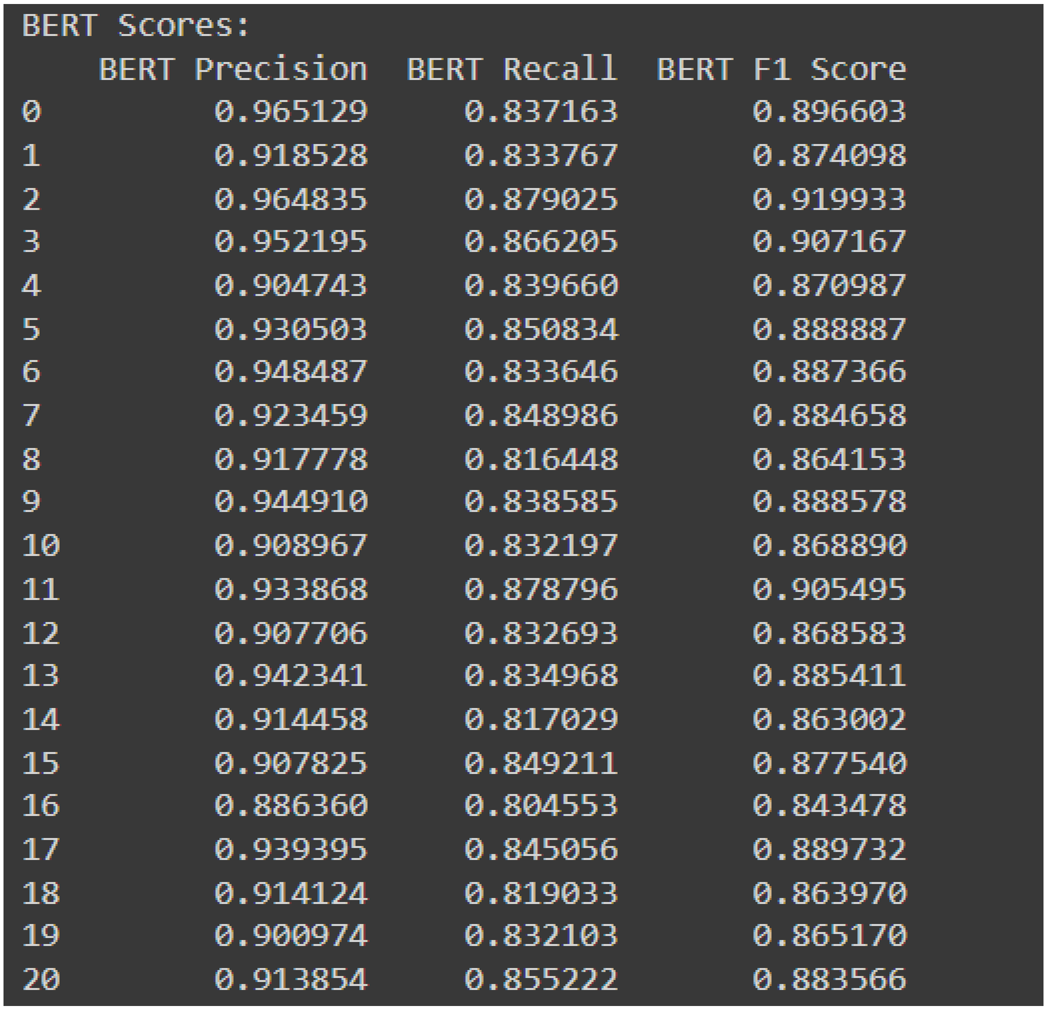
Evaluation Metrics:
Rogue Score: Measures of congruence exist between the predicted and the reference summaries.



```
    ROUGE-2 Recall  ROUGE-2 F1  ROUGE-L Precision  ROUGE-L Recall  ROUGE-L F1
0         1.000000    0.072539           0.038105        1.000000    0.073412
1         0.936508    0.109462           0.062008        0.984375    0.116667
2         1.000000    0.182289           0.100797        1.000000    0.183135
3         0.993151    0.237316           0.136490        1.000000    0.240196
4         0.959016    0.220755           0.130990        1.000000    0.231638
5         0.991228    0.120598           0.065304        1.000000    0.122601
6         0.990291    0.148041           0.081505        1.000000    0.150725
7         0.994845    0.236230           0.135323        1.000000    0.238386
8         1.000000    0.089552           0.047495        1.000000    0.090683
9         0.992308    0.118839           0.064153        1.000000    0.120571
10        0.972222    0.200957           0.116205        1.000000    0.208214
11        1.000000    0.179551           0.099124        1.000000    0.180369
12        0.990826    0.156635           0.086546        1.000000    0.159305
13        0.977011    0.470914           0.320000        1.000000    0.484848
14        0.986842    0.070192           0.037342        1.000000    0.071996
15        0.973214    0.101537           0.055501        1.000000    0.105165
16        0.976744    0.077778           0.041928        1.000000    0.080481
17        1.000000    0.089236           0.047161        1.000000    0.090074
18        0.971014    0.062881           0.033446        0.985714    0.064698
19        0.974359    0.071261           0.038424        1.000000    0.074005
20        0.984733    0.118349           0.064390        1.000000    0.120990
```

Fig: 8.1 Rogue score of first 20 rows

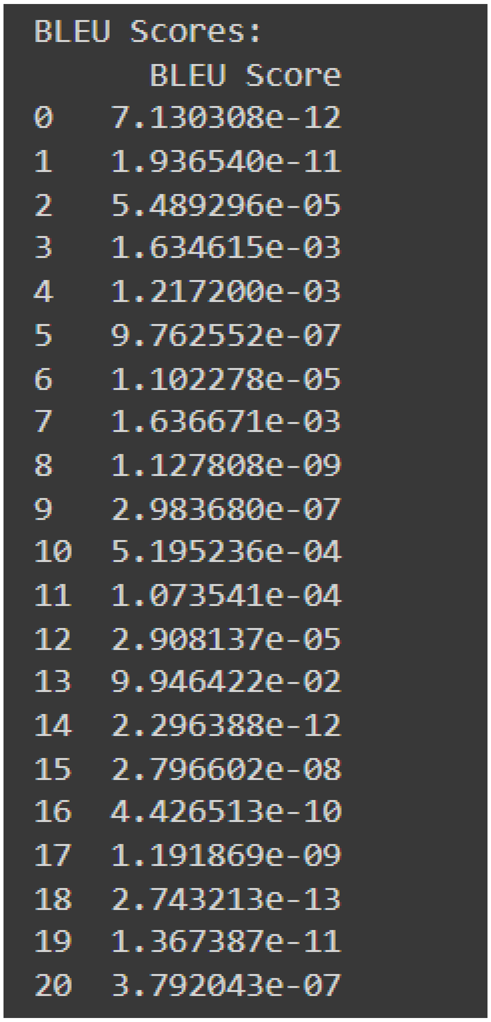
BERT Score: Measures similarity by computing BERT embeddings of predicted and reference texts.



```
BERT Scores:
    BERT Precision  BERT Recall  BERT F1 Score
0         0.965129     0.837163       0.896603
1         0.918528     0.833767       0.874098
2         0.964835     0.879025       0.919933
3         0.952195     0.866205       0.907167
4         0.904743     0.839660       0.870987
5         0.930503     0.850834       0.888887
6         0.948487     0.833646       0.887366
7         0.923459     0.848986       0.884658
8         0.917778     0.816448       0.864153
9         0.944910     0.838585       0.888578
10        0.908967     0.832197       0.868890
11        0.933868     0.878796       0.905495
12        0.907706     0.832693       0.868583
13        0.942341     0.834968       0.885411
14        0.914458     0.817029       0.863002
15        0.907825     0.849211       0.877540
16        0.886360     0.804553       0.843478
17        0.939395     0.845056       0.889732
18        0.914124     0.819033       0.863970
19        0.900974     0.832103       0.865170
20        0.913854     0.855222       0.883566
```

Fig: 8.2 BERT score of first 20 rows

BLEU Score: Assesses the level of precision by comparing n-grams of predicted text with reference text.



```
BLEU Scores:
      BLEU Score
0   7.130308e-12
1   1.936540e-11
2   5.489296e-05
3   1.634615e-03
4   1.217200e-03
5   9.762552e-07
6   1.102278e-05
7   1.636671e-03
8   1.127808e-09
9   2.983680e-07
10  5.195236e-04
11  1.073541e-04
12  2.908137e-05
13  9.946422e-02
14  2.296388e-12
15  2.796602e-08
16  4.426513e-10
17  1.191869e-09
18  2.743213e-13
19  1.367387e-11
20  3.792043e-07
```

Fig: 8.3 BLEU score of first 20 rows

Cross-Validation: Makes robust and generalizable the model recommendations.

The analysis of what these scores indicate is as follows:

ROUGE Scores

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap between the generated text and a reference text, focusing on recall.

- **ROUGE-1 (Unigram Overlap):** Measures the overlap of single words between the generated and reference texts.
- **ROUGE-2 (Bigram Overlap):** Measures the overlap of two-word sequences.
- **ROUGE-L (Longest Common Subsequence):** Considers the longest matching sequence of words to capture sentence-level structure.

From the scores:

- **High recall (1.0):** Indicates the generated text covers all the reference text content.
- **Low precision:** Implies the generated text includes many words not in the reference, suggesting verbosity or irrelevant content.

- **F-measure:** Balances precision and recall. While your recall is perfect, the low precision results in moderate F-measure scores.

BLEU Scores

BLEU (Bilingual Evaluation Understudy) measures how many n-grams in the candidate text match the n-grams in the reference text.

- **Scores near zero (e.g., 7.13×10−127.13 \times 10^{-12}7.13×10−12 to 0.03256):** Indicate very few n-gram overlaps, suggesting that the generated text differs significantly from the reference in terms of word sequences.
- **Higher scores (e.g., 0.09946422086659633):** Still quite low, suggesting minimal n-gram matches even when slightly better than others.

BERTScores

BERTScore uses BERT embeddings to measure the similarity between the generated and reference texts at a semantic level.

- **High precision (e.g., 0.9651), recall (e.g., 0.8372), and F1 scores (e.g., 0.8966):** Indicate that the generated text is semantically similar to the reference text, capturing the meaning well even if the exact wording differs.
- **Variation in scores:** Shows the consistency of the model in capturing the semantic essence across different outputs.

Analysis

1. **ROUGE:**
    - Perfect recall with low precision indicates that your model captures all necessary content but is verbose.
    - Moderate F-measure scores highlight the need to improve precision while maintaining high recall.
2. **BLEU:**
    - Low scores suggest that the generated text diverges significantly from the reference in exact wording and phrase structure.
    - This might indicate issues with fluency or that the generated text is too verbose or disjointed compared to the reference.
3. **BERTScores:**
    - High scores across the board indicate that despite the issues with exact matches (as seen in BLEU and precision in ROUGE), the generated text maintains high semantic similarity to the reference.

Recommendations for Improvement

- **Reduce Verbosity:** Work on improving the precision of the model to avoid unnecessary or irrelevant content. This could be achieved through better training data or fine-tuning techniques that emphasize brevity.
- **Fluency and Structure:** Since BLEU scores are low, focus on improving the grammaticality and coherence of the generated text to better match the reference text structures.
- **Balance Content and Form:** While high semantic similarity is achieved (as indicated by BERTScores), strive for a better balance between content coverage and textual precision to improve overall readability and utility.

By addressing these areas, you can enhance the overall performance of your text generation or summarization system, achieving better precision and maintaining high recall and semantic similarity.

Chapter 9: Challenges & Limitations

9.1 Challenges

- **Limited Domain-Specific Training**
  Current Large Language Models (LLMs) lack training on domain-specific data, resulting in inaccurate results for domain-specific retrievals and question-answering tasks.

- **Data Timeliness and Retraining Costs**
  LLMs do not possess up-to-date data, and the cost and time required for retraining or fine-tuning LLMs with the latest data can be prohibitive. Adjusting weights during this process may lead to significant changes in the model's output.

- **Managing Vast Volume of Patent Publications**
  The immense and rapidly growing volume of patent publications necessitates effective data storage with compression techniques to minimize information loss. Efficient retrieval techniques are crucial for accessing relevant information within this vast dataset.

- **Dynamic Data Management**
  Storing and managing large volumes of static data that require regular updates presents a challenge. Efficient searching within cost limits and low latency is essential.

- **Technical Challenges**
  - Local processing and text cleaning on the entire dataset became unfeasible due to its size.
  - Encountered limitations running the H2 database as Java couldn't be downloaded due to Colab's notebook instance setup.
  - Text cleaning algorithms significantly altered scientific terms, causing semantic loss and impacting the summarization ability of RAG (Retrieval-Augmented Generation).
  - Managing the trade-off between preserving semantics and reducing redundancy in sliding window chunking was challenging.
  - Using pre-trained RAG models for text summarization requires tweaking hyperparameters to effectively suit the task.

9.2 Limitations

- Pretrained text-to-embedding models tailored to our use case lacked comprehensive documentation, hindering effective implementation

- Hardware limitations exacerbate challenges in implementing pretrained text-to-embedding models due to increased computational demands for fine-tuning and inference
- Lack Of Faster Machines GPUs and Storage Capacity To Get Quicker Results.

Chapter 10: Conclusion/Recommendations

10.1 Recommendations & Business Impact

1. **Enhanced Accuracy in Patent Research**: Utilizing LLMs improves the precision and exhaustiveness of cited patent recommendations, ensuring comprehensive results and reducing the risk of overlooking vital patents and prior art.
2. **Significant Time Efficiency**: Automation reduces processing time from 4-8 weeks to just 1 day, saving approximately 96.25% of hours, thereby accelerating the research process and enabling quicker decision-making.
3. **Cost Savings**: Automation significantly lowers labor expenses by 50-70% and training costs by 30-50%, offering substantial financial benefits to the organization.
4. **Improved Resource Allocation**: By reducing the time and expertise required for manual patent searches, resources can be reallocated to other critical areas, fostering innovation and development within the organization.
5. **Reduction in Legal Risks**: Enhanced accuracy and comprehensive search results minimize the risk of errors and legal complications, providing a more reliable foundation for patent-related decisions and reducing potential litigations.
6. **Scalability and Maintenance**: The development of an easy-to-use, scalable, and easy-to-maintain model ensures long-term viability and adaptability, facilitating continuous improvement and expansion of the patent research process.
7. **Positive Economic Performance**: Compared to traditional methods, the automated approach aims to achieve a financial performance that meets or exceeds benchmarks, contributing to the overall economic health and competitive edge of the organization.

10.2 User Value from this project

We've successfully completed the code and verified its functionality using a representative sample from the population dataset. Due to constraints in digital infrastructure, we've operated with a scaled-down dataset. However, we're confident that with minor adjustments, the solution will seamlessly extend to the entire dataset. Our aim has been to deliver the best possible outcome within the existing parameters, ensuring that our solution is robust and adaptable to future scalability requirements.

Chapter 11: References

1. Patent Searching, Retrieval and Analysis: Project Paper Submitted by: Ms Misra S. Digital Library Home: Patent Searching, Retrieval and Analysis (juit.ac.in).

2. Basics of Patent Searching: By: Nigel S. Clarke:
   (PDF) The basics of patent searching (researchgate.net).
3. THE CHRONICLES OF RAG: THE RETRIEVER, THE CHUNK AND THE GENERATOR
4. HiQA A Hierarchical Contextual Augmentation RAG for Massive Documents QA
5. Embedding Compression in Recommender Systems: A Survey
6. Development and Testing of Retrieval Augmented Generation in Large Language Models - A Case Study Report
7. Retrieval Augmented Generation(RAG) using LlamaIndex and Mistral 7B   by Netra Prasad Neupane   Jan, 2024   Medium
8. Retrieval-Augmented Generation for Large Language Models-A Survey
9. Seven Failure Points When Engineering a Retrieval Augmented Generation System
10. The-state-of-the-art-on-Intellectual-Property-Analytics--IPA--- 2018 World-P

Thank You!!!