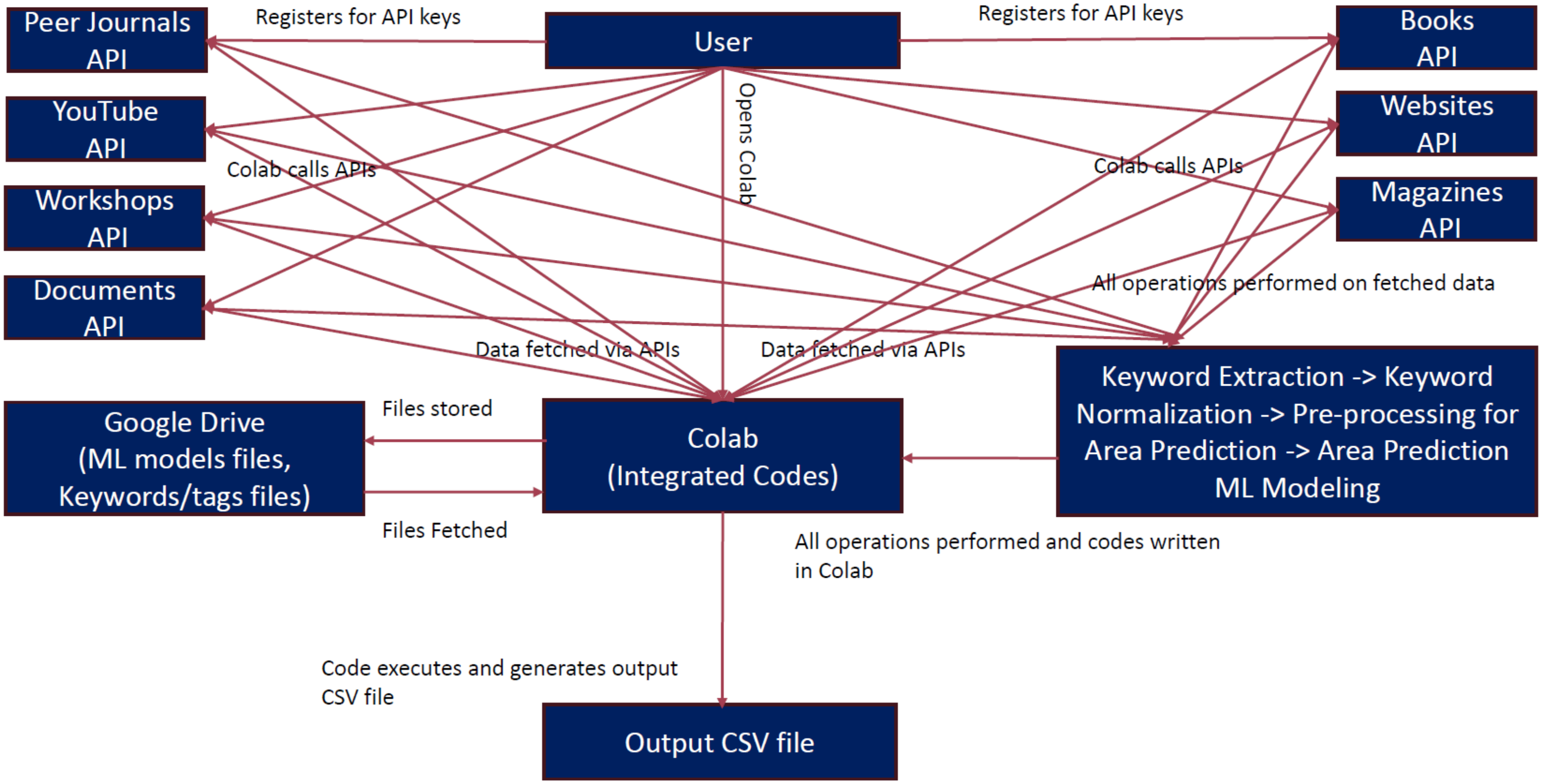
SYSTEM ARCHITECTURE


Peer Journals API
YouTube API
Workshops API
Documents API
Registers for API keys
User
Registers for API keys
Books API
Websites API
Magazines API
Opens Colab
Colab calls APIs
Colab calls APIs
All operations performed on fetched data
Data fetched via APIs
Data fetched via APIs
Google Drive (ML models files, Keywords/tags files)
Files stored
Files Fetched
Colab (Integrated Codes)
Keyword Extraction -> Keyword Normalization -> Pre-processing for Area Prediction -> Area Prediction ML Modeling
All operations performed and codes written in Colab
Code executes and generates output CSV file
Output CSV file

The following architecture outlines the workflow for an AI/ML-driven system designed to automate the gathering, processing, and categorization of research data in the field of continuous manufacturing. This system efficiently integrates data from multiple sources, applies advanced machine learning techniques, and ensures that the research is both accurate and relevant to the pharmaceutical manufacturing industry.

1. User Interaction:

- The **User** initiates the process by registering for **API keys** across multiple data sources, such as **Peer Journals**, **YouTube**, **Workshops**, **Books**, **Magazines**, **Websites**, and other relevant outlets. This step is essential for accessing the necessary research data.
- Upon receiving the API keys, the **User** accesses **Google Colab**, an integrated coding environment, to run the operations and execute the processes.

2. Data Collection:

- **Colab** interacts with the APIs to fetch metadata and content from 16 distinct data sources. These sources encompass a variety of materials, including research articles, documents, books, and videos related to continuous manufacturing.
- The collected data includes important metadata (such as titles, authors, abstracts, keywords) along with the content itself.

3. Keyword/Tag List for Filtering:

- A dynamic list of **keywords/tags** is maintained and stored in **Google Drive**. This list plays a crucial role in filtering the metadata collected by the APIs.
- The keywords, which are related to pharmaceutical continuous manufacturing (such as "API," "Oral Solid Dosage," "Pharmaceutical Processes"), ensure that only relevant research is included. The list can be manually updated, allowing the system to remain adaptable to emerging trends and new topics within the field.

4. Data Preprocessing and Area Prediction:

- The data undergoes a series of processing steps in **Google Colab**, where **keyword extraction** and **keyword normalization** are performed to further refine the data.
- **Pre-processing for area prediction** involves transforming the data to create dummy variables for each specific manufacturing area (e.g., "API," "Oral Solid Dosage"). This prepares the data for accurate classification using machine learning models.
- Embedding models, such as **BioBERT** and **SBERT**, are employed to generate meaningful representations of the data, which are then used for the prediction of relevant categories.

5. Machine Learning Model Training:

- After preprocessing, machine learning models, such as **Random Forest** and **Logistic Regression**, are trained on the processed data to predict the appropriate labels for each research document.
- These models are trained separately for each area label (e.g., "API," "Oral Solid Dosage") using the embeddings as features. Techniques like **SMOTE** are applied to address label imbalance, ensuring optimized model performance.

6. Data Storage (Final Model Files):

- After the machine learning models are trained and evaluated, the final **model files** (along with other necessary files such as keywords and tag lists) are stored in **Google Drive** for future use and deployment.
- The storage of these files ensures that the models can be reused or updated in the future without the need to retrain them from scratch, streamlining ongoing operations and research gathering.

7. Final Output:

- Upon successful model training and classification, the system generates an **output CSV file**, which contains the categorized research data. This file is structured to make the data easily accessible for further analysis, reporting, or integration into other systems.

Summary of Workflow:

- **User registers API keys** for data sources.
- **Colab** fetches research data via the APIs.
- **Google Drive** stores the dynamic **keywords/tags list** used to filter data based on relevance to pharmaceutical continuous manufacturing.
- The data is preprocessed, categorized, and labeled using machine learning models.
- **Google Drive** is also used to store the final **model files** and supporting resources.
- The system outputs a **CSV file** containing the categorized data.

Conclusion:

This system provides an automated, efficient, and scalable solution for gathering and categorizing research data in the continuously evolving field of pharmaceutical continuous manufacturing. By leveraging AI/ML models, the system ensures that only the most relevant and recent research is collected, categorized, and made available in a structured format. The integration of dynamic filtering through keywords and the use of advanced machine learning techniques enhances both the accuracy and efficiency of research gathering, benefiting stakeholders by ensuring timely access to high-quality, relevant research.

This workflow allows for continuous learning and scalability, ensuring that the system remains adaptable as new research emerges and as the field of continuous manufacturing evolves.