Project Documentation: Pharmaceutical Tender Data Processing and Analysis

Language: English

Project Overview:

This project focuses on processing pharmaceutical tender data extracted from bulk downloads. The primary goal is to standardize and enrich this data by extracting dosage information from various sources (e.g., supplier specifications) and associating it with the relevant authors and their affiliations, particularly their country of origin. This enriched data is then used to identify key contributors and their geographic distribution within the pharmaceutical tender landscape. The final output is a processed dataset suitable for further analysis, reporting, and potentially visualization.

Key Features:

* **Data Extraction:** The code extracts relevant data from an Excel file containing pharmaceutical tender information.
* **Dosage Extraction:** It utilizes regular expressions to extract dosage information (e.g., mg, %, UI) from supplier specifications, handling variations in formatting and units. Multiple approaches are implemented for dosage extraction from both suppliers and pharma.
* **Duplicate Removal:** Normalization techniques are applied to remove duplicate dosage entries.
* **OpenAlex Integration:** The code leverages the OpenAlex API to retrieve author details (name and country) based on their OpenAlex IDs.
* **Pairwise Author Analysis:** The processed data is structured to facilitate the analysis of collaborations between authors, specifically identifying author pairs and their associated countries.
* **Data Export:** The final processed data is exported to an Excel file and a CSV file.

Architecture and Design:

The project employs a modular design, with distinct functions responsible for specific tasks:

1. **Data Loading and Preprocessing:** Initial data loading from the Excel file.
2. **Dosage Extraction Functions:** Separate functions (`extract_dosage`, `remove_duplicates`) are defined for extracting and cleaning dosage information from supplier specifications. These functions handle diverse dosage formats and units.
3. **Author Country Lookup:** A function `get_author_country` utilizes the OpenAlex API to retrieve

author country information based on their OpenAlex ID.
4. **Pairwise Author Processing:** A function `create_author_pairs` generates all possible pairs of authors from the dataset and their corresponding countries.
5. **Data Combination and Transformation:** The extracted dosage information and author country data are combined into a single dataframe.
6. **Data Export:** The final processed dataframe is exported to the specified output formats.

The project utilizes Pandas for data manipulation and analysis, and the `requests` library for interacting with the OpenAlex API. Regular expressions (the `re` module) are crucial for dosage extraction.

Workflow and Logic:

1. **Data Loading:** The script begins by loading data from an Excel file containing pharmaceutical tender details.
2. **Dosage Extraction:** Dosage information is extracted from the 'EspecificacionComprador' and 'EspecificacionProveedor' columns using the `extract_dosage` function. The `remove_duplicates` function ensures that only unique dosage entries are retained. The dosage is then added to the dataframe as a new column.
3. **OpenAlex Lookup:** For each author ID in the dataset, the `get_author_country` function is called to retrieve the author's country of origin using the OpenAlex API.
4. **Pair Generation:** The `create_author_pairs` function generates all possible combinations of authors (including both author A-author B and author B-author A).
5. **Data Merging:** The extracted dosage information and author country data are combined into a new dataframe using the author IDs as the key.
6. **Output:** The final processed dataframe is exported to an Excel and CSV file.

Key Concepts and Techniques:

* **Regular Expressions:** Used for robust dosage extraction, accommodating various formats and units.
* **Pandas DataFrames:** For data manipulation, cleaning, and transformation.
* **OpenAlex API:** Used to retrieve author details (name and country) programmatically.
* **Itertools (permutations):** Used to generate all possible pairs of authors from the dataset.
* **String Manipulation:** Used for normalizing dosage values and removing whitespace.
* **API Calls:** Using `requests` to communicate with the OpenAlex API.

Error Handling and Performance:

* **Missing Values:** The code handles missing values in the 'authorships.author.id' column by skipping rows that do not have a valid author ID.
* **Invalid OpenAlex IDs:** The `get_author_country` function includes error handling to deal with invalid OpenAlex IDs, returning "Invalid ID" in such cases.
* **API Rate Limiting:** While not explicitly implemented, production use should include error handling for potential API rate limits and consider implementing retry mechanisms.

* **Performance:** The performance of the dosage extraction function could be further optimized by exploring more efficient regular expression patterns or employing pre-compiled regular expressions. The API calls can be time consuming, and caching could be beneficial for frequently accessed OpenAlex IDs.

Potential Challenges and Considerations:

* **Inconsistent Dosage Formats:** Variations in dosage formats across different suppliers can be challenging to handle effectively. The current regex may require refinement to cover a wider range of formats.
* **API Rate Limits:** The OpenAlex API may impose rate limits on the number of requests. Consider implementing error handling and retry mechanisms to handle these limits.
* **Data Quality:** The quality of the input data (Excel file) is crucial. Inconsistent formatting or errors in the source data can negatively impact the accuracy of the extracted information.
* **Scalability:** For very large datasets, the performance of the OpenAlex API calls could become a bottleneck. Consider implementing batch processing or using a more scalable API solution.

Future Enhancements:

* **More Robust Dosage Extraction:** Refine the regular expressions to handle a wider range of dosage formats and units. Consider using machine learning techniques to improve the accuracy of dosage extraction.
* **Error Logging and Reporting:** Implement a comprehensive error logging mechanism to track and diagnose issues during data processing.
* **Data Validation:** Add data validation steps to ensure the accuracy and consistency of the extracted data.
* **Visualization:** Create visualizations (e.g., maps, charts) to explore the geographic distribution of authors and their collaborations.
* **Caching:** Implement caching mechanisms to reduce the number of API calls to the OpenAlex API.
* **User Interface:** Develop a user interface to allow users to upload data and configure processing parameters.

Summary:

This project provides a framework for extracting, standardizing, and enriching pharmaceutical tender data by integrating dosage information and author affiliations. The code utilizes regular expressions, the OpenAlex API, and Pandas to achieve its goals. While the current implementation addresses the core requirements, future enhancements could further improve accuracy, robustness, and scalability. Proper maintenance and support will be crucial for ensuring the continued value of this data processing pipeline.