Project Documentation: OpenAlex Author Country Extraction

Version: 1.0
Date: October 26, 2023
Author: AI Technical Writer

1. Project Overview

This project focuses on extracting and enriching author affiliation data from a dataset of publications. Specifically, it leverages the OpenAlex API to determine the country of affiliation for each author listed in the dataset. The primary goal is to supplement the existing dataset with geographical information, facilitating further analysis and potentially improving the context of the research represented within the data. Key features include:

* **OpenAlex API Integration:** Utilizing the OpenAlex API to retrieve author details and affiliation information.
* **Data Cleaning & Filtering:** Selecting relevant columns from the input dataset and removing irrelevant data.
* **Geographic Enrichment:** Mapping authors to their country of affiliation using the OpenAlex API.
* **Data Transformation:** Creating new columns containing author pairs and their corresponding country affiliations.

2. Architecture and Design

The project employs a modular architecture centered around Python scripting, leveraging the `pandas` library for data manipulation and the `requests` library for interacting with the OpenAlex API. The core logic is encapsulated within functions, promoting reusability and maintainability.

The architecture can be broadly divided into the following components:

1. **Data Input:** The project begins by loading a CSV file containing publication data.
2. **Data Preprocessing:** The DataFrame is filtered to retain only the necessary columns related to author IDs and country affiliations.
3. **API Interaction:** The `get_country_from_openalex` and `get_author_country` functions are called to retrieve country data from the OpenAlex API, based on the author IDs present in the DataFrame.
4. **Data Transformation:** The extracted country information is used to generate new columns representing author pairs and their affiliated countries.
5. **Output:** The enriched DataFrame is saved to a new CSV and Excel file.

The logic is primarily driven by function calls and vectorized operations within `pandas`, prioritizing efficiency and scalability. The use of functions helps to isolate concerns and improve code readability.

3. Key Functionalities

* **OpenAlex Author Country Lookup:** The core functionality is the ability to determine the country of affiliation for a given OpenAlex author ID using the OpenAlex API. This function handles API requests, parses the JSON response, and extracts the relevant country code.
* **Author Pair Generation:** The project generates pairs of authors from the dataset and identifies their respective countries of affiliation. This allows for an examination of collaborative efforts and geographic distribution of researchers.
* **Data Enrichment:** The process of adding country affiliation data to the original dataset is crucial for enhanced analytical capabilities.
* **Data Transformation and Cleaning:** The code filters, cleans, and transforms data to ensure that it is in the correct format for analysis. It handles missing data gracefully.

4. Workflow and Logic

1. **Data Loading:** The script begins by reading the CSV file containing the publication data.
2. **Initial Filtering:** The DataFrame is filtered to retain only the relevant columns: `id`, `type`, `type_crossref`, `cited_by_count`, `referenced_works_count`, `publication_year`, `authorships.author.id`, and `authorships.countries`.
3. **API Call for Individual Authors:** For each author ID in the filtered DataFrame, the `get_country_from_openalex` function is called. This function makes an API request to OpenAlex, retrieves the author's details, and parses the response to extract the country code.
4. **Handling of Missing/Invalid Data:** If an API call fails or returns an invalid response, the function returns "Invalid ID," which is then handled appropriately.
5. **Pair Generation Logic:** The `create_author_pairs` function generates all possible combinations of authors from the dataset, along with their associated countries.
6. **Output Generation:** The enriched DataFrame is saved to a new CSV and Excel file, capturing the original data along with the added country affiliation information.

5. Key Concepts and Techniques

* **API Integration:** The project utilizes the `requests` library to make HTTP requests to the OpenAlex API. Understanding API authentication, request parameters, and JSON parsing is essential.
* **Pandas Data Manipulation:** The `pandas` library is extensively used for data loading, filtering, cleaning, and transformation.
* **List Comprehensions:** List comprehensions are used for efficient data processing and transformation.

* **Itertools:** The `itertools` library is used to generate combinations and permutations.
* **Error Handling:** The code includes basic error handling to gracefully manage API failures and invalid data. Specifically, it checks for `None` values and handles potential errors from the OpenAlex API.
* **String Manipulation:** The code uses string manipulation techniques to extract author IDs from URLs.

6. Error Handling and Performance

* **API Errors:** The `get_country_from_openalex` function includes error handling to catch potential API errors (e.g., rate limiting, network issues) and returns "Invalid ID" in case of failure.
* **Missing Data:** The `create_author_pairs` function handles missing author IDs by skipping those authors.
* **Performance:** Vectorized operations in `pandas` are utilized to optimize performance. Large datasets may require optimization techniques such as chunking or using more efficient data structures.

7. Potential Challenges and Considerations

* **OpenAlex API Rate Limits:** The OpenAlex API has rate limits. The script needs to be designed to handle these limits gracefully (e.g., by implementing delays between API calls).
* **API Changes:** The OpenAlex API may change over time, requiring updates to the script.
* **Data Quality:** The accuracy of the country affiliation data depends on the quality of the data in the OpenAlex API.
* **Large Datasets:** Processing very large datasets may require significant computational resources and optimization.
* **Author ID Uniqueness:** Ensuring that OpenAlex IDs are unique across the dataset is crucial for accurate results.

8. Future Enhancements

* **Caching:** Implement caching of OpenAlex API responses to reduce the number of API calls and improve performance.
* **Error Logging:** Implement more robust error logging to track API failures and data quality issues.
* **Database Integration:** Store the enriched data in a database for easier querying and analysis.
* **Geographic Visualization:** Visualize the distribution of authors by country using a mapping tool.
* **Advanced Filtering:** Add more sophisticated filtering options based on author criteria (e.g., publication year, research area).
* **Bidirectional Matching**: Implementing a mechanism to find authors who have collaborated with the same country of origin.

9. Summary

This project successfully extracts and enriches publication data with country affiliation information using the OpenAlex API. The code is modular, well-documented, and designed to handle potential errors and limitations. The resulting enriched dataset will be valuable for various research and analysis tasks. Maintenance will require monitoring the OpenAlex API for changes and addressing any errors or performance issues that may arise. The primary impact of this project is the addition of a critical geographic dimension to the existing publication data, enabling more sophisticated and insightful analyses.