Report: PubMed Paper Fetcher and Filter Tool

Objective

The objective of this project is to develop a Python-based tool that fetches PubMed articles using a user-defined query, filters them based on the affiliation of authors, and outputs the results in a structured CSV format. The tool aims to identify non-academic authors and highlight affiliations with pharmaceutical or biotechnology companies.

Approach

1. Data Retrieval

- **API Integration**: The tool utilizes the Entrez Programming Utilities (E-utilities) from NCBI to interact with the PubMed database.
- **Search Process**:
    - The `esearch` endpoint retrieves PubMed IDs (`PMID`) for articles based on a user-defined query.
    - The `efetch` endpoint fetches detailed metadata for articles corresponding to the retrieved PMIDs.
- **Query Flexibility**: Users input a query string (e.g., keywords, author names, or journal titles) to fetch articles relevant to their interests.

2. Filtering Process

- **Author Categorization**:
    - Authors are categorized based on their affiliations, extracted from the metadata.
    - Non-academic authors are identified by checking if their affiliations lack academic keywords (e.g., “university,” “college”).
- **Company Affiliation Identification**:
    - Affiliations are further analyzed for pharmaceutical or biotechnology keywords (e.g., “pharma,” “biotech”).
- **Data Extraction**:
    - For each article, the following information is extracted:
        - **PMID**
        - **Title**
        - **Publication Date**
        - **Non-academic Authors**
        - **Company Affiliations**
        - **Corresponding Author Email**

3. Output

- **CSV Report**:
  - Results are written into a CSV file or displayed on the console.
  - The output includes both author and article-level details for easy analysis.
- **Debug Mode**: Provides intermediate data points during execution for better traceability and validation.

Methodology

1. Affiliation Analysis

- Keywords for academic institutions and companies are used to classify affiliations.
- Parsing and natural language techniques identify patterns within the affiliation text.

2. Publication Date Standardization

- Metadata for publication dates is processed to provide a uniform date format (`YYYY-MM-DD`), even for incomplete data (e.g., year-only).

3. Corresponding Author Email Extraction

- Email addresses are extracted directly from the affiliation text, leveraging regex-like pattern matching for tokens containing "@."

Results

Expected Outcomes

1. **Filtered Articles**:
   - The tool effectively narrows down articles with at least one non-academic author.
   - Highlights company affiliations relevant to the pharmaceutical or biotechnology sectors.
2. **Actionable Data**:
   - Corresponding author emails enable direct communication with industry authors.
   - Data provides insights into industry-academia collaborations.
3. **Scalability**:
   - Flexible query handling allows the tool to adapt to various research interests.

Sample Output

PubMed ID	Title	Publication Date	Non-academic Authors	Company Affiliations	Email
12345678	Advancements in AI	2024-01-01	John Doe	XYZ Biotech	john@xyz.com
23456789	Drug Development Trends	2023-12-15	Jane Smith	ABC Pharmaceuticals	jane@abcpharma.com

Challenges and Limitations

1. **Data Quality**:
    - The completeness and accuracy of metadata in PubMed records directly impact the tool's effectiveness.
2. **Keyword-based Filtering**:
    - False positives/negatives may occur if affiliations contain ambiguous terms.
3. **Limited to PubMed**:
    - The tool focuses solely on PubMed data and does not fetch articles from other repositories.

Conclusion

The PubMed Paper Fetcher and Filter Tool is an efficient utility for researchers, analysts, and industry professionals to analyse PubMed articles. By automating the fetching, filtering, and categorization of articles, the tool saves time, provides valuable insights, and facilitates industry-specific research.

Future enhancements could include integrating machine learning models to improve affiliation classification and expanding support for other research databases.