**Report: PubMed Research Paper Extraction and Filtering System**

**1. Introduction**

The objective of this project is to extract research papers from PubMed based on a user-specified query, filter papers with at least one author affiliated with a pharmaceutical or biotech company, and export the results in a structured CSV format. The system is designed to ensure efficient data retrieval, modularity, and scalability, making it suitable for future enhancements.

**2. Approach**

The approach involves developing a Python-based tool that interacts with the PubMed API to retrieve research articles, processes the author data to identify industry-affiliated contributors, and exports the refined dataset into a CSV file. The project follows best practices in software development, including:

* **Typed Python** for improved code readability and reliability.
* **Poetry for dependency management** to ensure reproducibility.
* **Command-line interface (CLI) implementation** for user-friendly interaction.
* **Structured data handling** using dictionaries and lists.
* **Test-driven development (TDD)** principles where applicable.

**3. Methodology**

**3.1 Data Retrieval from PubMed**

1. The program queries PubMed using user-defined search terms.
2. Results are retrieved in JSON format, including metadata such as PubMed ID, Title, Authors, Affiliations, and Publication Date.

**3.2 Filtering Industry-Affiliated Authors**

1. Authors’ affiliations are parsed to detect pharmaceutical or biotech company associations.
2. Authors with non-academic affiliations are extracted and labeled separately.

**3.3 Data Processing and Formatting**

1. Author names are formatted into a semicolon-separated string for CSV output.
2. Additional fields such as Corresponding Author Email and Company Affiliation(s) are extracted.
3. Missing data fields are replaced with "N/A" to maintain consistency.

**3.4 Exporting Data to CSV**

1. The processed data is structured into a well-defined CSV format with the following fields:
   * **PubmedID**
   * **Title**
   * **PublicationDate**
   * **Authors** (formatted as "Name1; Name2; Name3")
   * **Non-academic Author(s)**
   * **Company Affiliation(s)**
   * **Corresponding Author Email**
2. The CSV is generated with UTF-8 encoding to ensure compatibility across platforms.

**4. Results**

The refined system successfully extracts and filters research papers based on industry affiliations, outputting structured CSV files. Sample results include:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PubmedID** | **Title** | **PublicationDate** | **Authors** | **Non-academic Author(s)** | **Company Affiliation(s)** | **Corresponding Author Email** |
| 40090539 | Exploring the Intricacies of Antimicrobial Resistance | 2025 | Haq Shahbaz Ul; Ling Wang | Haq Shahbaz Ul; Gao Fenfei | Shantou University Medical College | ffgao@stu.edu.cn |
| 40090144 | Black soldier fly larvae mediate Zinc and Chromium transformation | 2025 | Xu Chao; Wang Panpan | Xu Chao | Huazhong Agricultural University | qxyuan@mail.hzau.edu.cn |
| 40090092 | Engineered magnetic particles derived from steelmaking dust | 2025 | Rokowska Anna; Dziewit Lukasz | Rokowska Anna | University of Warsaw | mskalny@agh.edu.pl |

**5. Conclusion and Future Improvements**

The system provides an automated, scalable solution for extracting and analyzing research articles from PubMed with a focus on industry affiliations. Future improvements may include:

* **Enhancing filtering algorithms** to improve accuracy in detecting industry affiliations.
* **Adding support for more file formats** such as JSON and Excel.
* **Integration with NLP models** to classify affiliations more precisely.
* **Deploying as a web-based service** for wider accessibility.

This tool serves as a foundation for further research analytics applications and demonstrates the effectiveness of structured data extraction in bibliometric analysis.

**✅ Final Outcome:**

The project successfully extracts, filters, and exports relevant PubMed data while maintaining modularity and efficiency.