Rishabh Khandelwal

PubMed Paper Scraper: Summary Report

1. Introduction

The PubMed Paper Scraper is a Python CLI tool designed to fetch research papers from the PubMed API based on user-defined queries. It also analyses author affiliations to identify non-academic authors, particularly those affiliated with pharmaceutical and biotech companies. The output is presented in a CSV format, providing users with valuable insights into research papers and their associated affiliations.

Key Features:

- **Full PubMed query syntax support** for precise paper retrieval.
- **Non-academic author detection** based on affiliation heuristics.
- **CSV output** with fields such as PubMed ID (PMID), title, publication date, author names, company affiliations, and email addresses.
- A **CLI interface** for easy interaction with the tool.

---

2. Approach

The development of this tool involved a combination of utilizing the PubMed API, incorporating LLM-based text classification, and leveraging heuristics to filter affiliations and detect pharmaceutical or biotech-related authors.

2.1. Data Collection

The first step was to interact with the **PubMed API** to fetch research papers based on a user-provided search query. The query allows for precise filtering of research papers, pulling relevant data including:

- **PMID** (PubMed Identifier)
- **Title** of the paper
- **Publication date**
- **Author list** and **affiliation** details
- **Corresponding author email** (if available)

2.2. Identifying Non-academic Authors

To identify authors affiliated with pharmaceutical or biotech companies, we used a two-pronged approach:

1. **Heuristic-based filtering**: Applied basic heuristics to detect keywords like "Inc", "Pharma", "Biotech", and "LLC" in affiliations and emails, flagging affiliations likely tied to industry sectors rather than academic institutions. For academic institutions, checking for keywords like “Institution”, “University”, etc.

2. **LLM-based classification (work in progress/partially implemented)**: Using a language model (LLama 3:8B via Ollama) - classified affiliations as academic or non-academic by evaluating the presence of industry-specific terms in the affiliation strings.

2.3. Tool Design

- The tool was built as a **Python CLI** using the argparse library for handling user inputs and options.
- **Poetry** was used for dependency management, ensuring a robust and reproducible environment for development and deployment.
- Output is saved as a **CSV file**, with relevant fields clearly labeled to provide meaningful insights to the user.

---

3. Methodology

3.1. Heuristic Filtering

In addition to the LLM classification, we implemented a **heuristic filter** that checked for keywords in the affiliation and email domains:

- **Pharmaceutical/biotech keywords**: "Pharma", "Biotech", "Inc", "Labs", "R&D".
- **Email domain checks**: Recognizing company email domains like @pharma.com, @biotech.org to detect industry-based affiliations.

3.2. CLI Design

The command-line interface (CLI) was designed to support the following key options:

- query: User-provided PubMed search query.
- --file: Path to save the output CSV.
- --debug: Enable detailed logging for debugging purposes.
- --help: Display CLI help and usage instructions.

---

4. Results

The tool was successfully developed and tested with the following outcomes:

- **PubMed Data Retrieval**: The tool was able to fetch relevant papers based on a variety of search queries, including keywords related to cancer, drug research, and immunotherapy.
- **Non-academic Author Identification**: The model correctly identified pharmaceutical and biotech affiliations in most test cases. A few edge cases (e.g., companies with academic-sounding names) were manually verified.
- **Output in CSV**: The results were formatted into a CSV file with the following fields:
    - **PubMedID** (PMID)

- **Title**
- **Publication Date**
- **Non-academic Author(s)**
- **Company Affiliation(s)**
- **Corresponding Author Email**

Pub Med ID	Title	Publication Dat	Authors	Company Affili	Corresponding
40217112	Privacy-preservir	2025-Apr-11	['N/A']	['N/A']	N/A
40216796	A multi-site, mul	2025-Apr-11	['Asante Ntata']	['N/A']	N/A
40216216	Electroencephalc	2025-Apr-09	['Saeideh Davou	['N/A']	saeideh.davoudi

---

5. Challenges

- **False Positives**: Some affiliations were misclassified, particularly companies with academic-sounding names or research institutions with industry partnerships.

---

6. Future Work

Future improvements could focus on:

- **Enhanced Filters**: Expanding the keyword list and considering context, using LLMs (partially implemented).
- **User Feedback Loop**: Allowing users to flag incorrect classifications, enabling continuous model improvement.
- **Performance Optimizations**: Handling larger datasets more efficiently by implementing caching and parallel processing.

---

7. Conclusion

The PubMed Paper Scraper provides an efficient, easy-to-use tool for identifying research papers related to pharmaceutical and biotech companies. By integrating classification with heuristic filtering, the tool can accurately identify non-academic authors and output structured results. With further improvements, it can be a valuable resource for research professionals looking to analyze scientific papers and author affiliations.