PubMed Fetcher: A Python CLI Tool for Retrieving Papers with Non-Academic Authors

Varshini D
varshinid.tech@gmail.com

July 12, 2025

1 Introduction

The PubMed Fetcher is a Python command-line tool designed to query the PubMed database, retrieve research papers with at least one author affiliated with a pharmaceutical or biotech company, and output the results in a CSV file. This report outlines the approach, methodology, and results of the project, demonstrating its functionality and implementation details.

2 Approach

The project adopts a modular design to fetch, filter, and output PubMed data. It leverages the PubMed API for data retrieval, applies a heuristic to identify non-academic authors, and formats results for CSV output. The tool is built with Python, using Poetry for dependency management, Biopython for API access, and Pandas for data handling. The implementation focuses on simplicity, reliability, and compliance with PubMed API requirements.

3 Methodology

3.1 Project Structure

The project is organized into modular components:

- `pubmed_fetcher/client.py`: Handles PubMed API queries and data parsing.
- `pubmed_fetcher/filter.py`: Filters papers with non-academic authors.
- `pubmed_fetcher/output.py`: Writes results to CSV or console.
- `scripts/cli.py`: Provides the command-line interface.
- `pyproject.toml`: Configures dependencies and CLI entry point.
- `README.md`: Documents setup and usage.

3.2 Implementation

The tool uses Biopython's `Entrez` module to query PubMed with user-provided search terms (e.g., "pfizer vaccine"). The `PubMedClient` class retrieves up to 10 papers, parsing XML data into structured records with fields: PubmedID, Title, Publication Date, Non-academic Author(s), Company Affiliation(s), and Corresponding Author Email. A heuristic identifies non-academic authors by checking affiliations for keywords like "pharma" or "biotech" while excluding academic terms like "university." The `AuthorFilter` class ensures only papers with non-academic authors are included, and the `OutputWriter` class generates CSV output using Pandas.

3.3 Challenges and Solutions

An initial `list index out of range` error in `client.py` occurred due to missing `AffiliationInfo` in some PubMed records. This was resolved by adding robust checks to handle empty or missing data. Git push conflicts were addressed by force-pushing to the repository after verifying local completeness. The heuristic's simplicity may miss some non-academic affiliations, but it effectively identifies clear cases (e.g., Pfizer, Moderna).

3.4 Dependencies

- **Python** ($\geq$3.13): Core language.
- **Biopython** (1.83): PubMed API access.
- **Pandas** (2.2.3): CSV output.
- **Poetry**: Dependency management and CLI setup.

4 Results

The tool was tested with the query "pfizer vaccine" using the command:

```
poetry run get-papers-list "pfizer vaccine" -f output.csv -d
```

The output CSV (`output.csv`) contained three papers with non-academic authors:

- **PubmedID 40640198**: Title: "Immunogenicity and protective efficacy of an intranasal neuraminidase-based influenza vaccine with bacterial cell membrane-derived adjuvants." Non-academic author: Zimmermann Joseph (Inspirevax Inc.).
- **PubmedID 40639194**: Title: "IxsS7: A novel biomarker for Ixodes scapularis tick bite exposure in humans." Non-academic authors: Kelly Patrick H, Stark James H (Pfizer Research & Development).
- **PubmedID 40639020**: Title: "Approaches to overcome the current treatment plateau in immunotherapy." Non-academic authors: Zheng Wei, Ibrahim Ramy (Moderna Inc., Georgiamune Inc.).

The CLI supports debug logging (`-d`) and help (`-help`). The repository is hosted at `https://github.com/Varshini-D777/pubmed-fetcher`.

5 Conclusion

The PubMed Fetcher successfully retrieves and filters PubMed papers with non-academic authors, meeting the project requirements. The modular design, robust error handling, and clear documentation ensure usability and maintainability. Future improvements could include refining the heuristic for broader affiliation detection.