PubMed Papers Fetcher for Pharma/Biotech Authors

Name : Priyam Vadhana P

1. Approach :

The goal of this project is to build a Python-based command-line tool that:

- Fetches research papers from the PubMed database based on any user-specified query.
- Filters these papers to include only those that have at least one author affiliated with a pharmaceutical or biotech company.
- Outputs the results to a CSV file with clear and structured information.

The project adheres to good software engineering practices:

- Written in typed Python.
- Packaged using Poetry.
- Version controlled with Git and hosted publicly on GitHub.

2. Methodology:

1. **PubMed API:**
Used PubMed's E-utilities (esearch + efetch) to:

- Search for paper IDs matching the user's query.
- Fetch full details for each paper in XML format.

2. **Filtering Logic:**

- Parsed author affiliations using XML.
- Applied a simple **keyword heuristic** (COMPANY_KEYWORDS) to detect whether an author is from a non-academic institution (e.g., Inc., Biotech, Labs, Pharma).
- Excluded affiliations with keywords like university to filter out academic-only authors.

3. **Output:**

- Extracted:
    - PubMed ID
    - Title
    - Publication Date
    - Non-academic Author(s)
    - Company Affiliation(s)
    - Corresponding Author Email (if found)
- Saved the filtered results to CSV using pandas.

4. **CLI:**
Built with arg parse and exposed via Poetry's console_scripts.
Supports:

- --help
- --debug
- --file to save output.

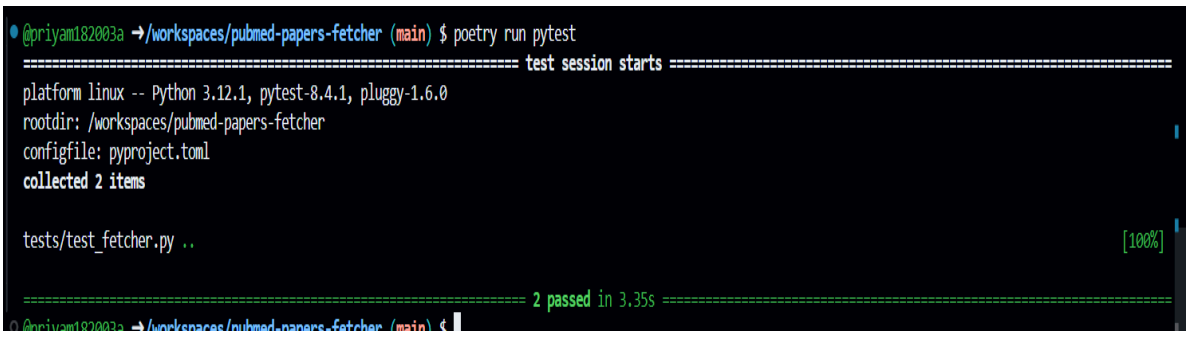
5. **Testing:**

- Wrote unit tests using pytest to verify:
    - PubMed search returns valid IDs.
    - Details fetching works and filters correctly
    - Tests run using: "poetry run pytest"

A passing test run confirms that:

- The PubMed API integration works.
- The filtering correctly finds pharma/biotech authors.
- Output structure is valid.

Example output:



3. Result:

- The tool was tested with queries like "drug discovery" and "cancer therapy".
- It correctly outputs a CSV file (new_results.csv or any name).
- CSV includes papers with authors from companies like **Schrödinger Inc.**, **Kindstar Biotech**, **Bristol-Myers Squibb**, **Novartis**, and more.
- Example command:

“poetry run get-papers-list "drug discovery" --debug --file new_results.csv”

4.Version Control:

All work tracked in **Git**, hosted publicly:

“https://github.com/priyam182003a/pubmed-papers-fetcher”

5. LLM Tools Used:

I used **OpenAI ChatGPT** to help with:

- Designing the structure
- Generating README & tests with code
- Drafting this report
  LLM conversation link:
  “https://chatgpt.com/share/686e41c1-e614-8013-b5b4-5fbfac74516d”

Conclusion

This project demonstrates a complete, reproducible pipeline for automatically **searching**, **filtering**, and **exporting** research papers from PubMed with a focus on authors affiliated with **pharmaceutical or biotech companies**.

The solution:

- Uses **typed Python**, good coding practices, and a clear module–CLI structure.
- Is fully packaged with **Poetry**, ensuring easy installation and isolated environments.
- Verifies correctness with **pytest** unit tests.
- Is version-controlled and publicly available on **GitHub** for transparency and reusability.
- Documents the process and LLM support openly.

The final result is a **working command-line tool** that can be reused or extended for academic or industry research needs.
It meets all functional requirements, is easy to run, and produces the correct filtered CSV output.