PubMed Research Paper Filter

Problem Statement :

Build a Python command-line tool that:

- Fetches research papers from PubMed using a user-provided query.
- Filters papers with at least one non-academic author affiliated with pharma/biotech companies.
- Exports the final filtered results into a CSV file.

Tools & Technologies Used :

Tool /Library	**Purpose /Usage**
Python 3.8+	Main Programming language
Requests	Send HTTP requests to the PubMed API
argparse	Parse command-line aruguments
csv	Write output data into csv files
Xml.etree.ElementTree	Parse XML responses from the PubMed API
PubMed E-Utilities API	Retrieve biomedical research metadata
Poetry(Optional)	Dependency & environment management
Command Line (CLI)	Run the script with various input aruguments
Git & Github	Version control and project hosting
VS Code / Replit	Code editor / online IDE for writing and testing code

Project Structure :

pubmed-project/

```
│
├── get_papers_list.py       # Main script with core logic
├── pyproject.toml           # Poetry dependency & CLI configuration
├── README.md                # Project documentation
└── results.csv              # Output file (auto-generated)
```

Key Features

✅ Accepts full PubMed query syntax

✅ Filters based on company affiliations using heuristics: pharma, biotech, labs, therapeutics

✅ Command-line options:

- -h / --help → Show help message
- -d / --debug → Enable verbose debug logs
- -f / --file → Specify output CSV filename

How to Run

Step 1: Setup

1. Install Visual Studio Code or use Replit.
2. Create a new folder (e.g., pubmed-project).
3. Inside it, create a file named get_papers_list.py and paste the core Python logic.
4. Open Terminal in VS Code and install the necessary package:

```
pip install requests
```

Step 2: Running the Script

Run the script using:

```
python get_papers_list.py "cancer research" -f cancer_results.csv -d
```

This command will fetch PubMed data for "cancer research", filter it, and store results in cancer_results.csv.

Step 3: Poetry Setup

1. Create a file pyproject.toml for dependency tracking.
2. Install Poetry if not available:

   python -m poetry add requests

   python -m poetry install

3. Run the script:
   poetry run python get_papers_list.py "cancer research" -f cancer_results.csv

Methodology / Logic

1. Fetch PubMed IDs using **esearch** endpoint.
2. For each ID, get metadata using **efetch**.
3. Parse **XML response** for author names and affiliations.
4. Filter based on company-related keywords:
   - pharma, biotech, labs, therapeutics, inc, ltd
5. Export relevant results to **CSV**.

Output Example :

PubmedID	**Title**	**Publication Date**	**Non-academic Author(s)**	**Company Affiliation (s)**	**Corresponding Author Email**
40654116	Boosting PDT with DPA-NI-Bu: high photocytotoxicity through redox homeostasi	2025	Wu	Key Laboratory for Tibet Plateau Phytochemistry of Qinghai Province, College of Pharmacy, Qinghai Minzu University, Xining 810007,	

	s perturbation.			Qinghai, P. R. China.	
40654114	Cost-effectiveness analysis of pembrolizumab as an adjuvant treatment of early-stage non-small cell lung cancer following complete resection and platinum-based chemotherapy in Canada.	2025	Wehler, Langevin, Bensimon, Leighl, Hu, Xu, Arunachalam, Insinga	Health Economic and Decision Sciences, Merck & Co., Inc., West Point, PA, USA. Health Economics Outcomes Research, Merck Canada Inc., Kirkland, QC, Canada. Health Economics and Outcomes Research, Analysis Group, Inc., Boston, MA, USA. Division of Medical Oncology, Princess Margaret Cancer Centre, Toronto, Canada. HTA Statistics, Merck & Co., Inc., West Point, PA, USA.	

	Work Title	Publication Date	Non-academic Author(s)	Company Affiliation(s)	Corresponding Author Email	
	Workflow studies and their interpretation.	2025	Cui, Sun, Yuan, Tian, Xie, Xu, Li	Department of Pharmacy, College of Veterinary Medicine, Sichuan Agricultural University, Chengdu, Sichuan 611130, China.		
40654116	Boosting PDT with DPA-NI-Bu: high photocytotoxicity through redox homeostasis perturbation.	2025	Wu	Key Laboratory of Tibetan Plateau Phytochemistry of Qinghai Province, College of Pharmacy, Qinghai Minzu University, Xining 810007, Qinghai, P. R. China.	—	

Contact

Made by:

Kalvapalli Venkata Srilatha
Email : kvslatha2003@gamil.com
Location: Gachibowli, Hyderabad