Aganitha Take Home Exercise - REPORT

PubMed Research Paper Filter CLI Tool

Submitted by: **Dharshini C**

Objective

To develop a command-line tool that fetches research papers from **PubMed** based on a user-provided query, filters for **non-academic authors** (typically affiliated with pharmaceutical or biotech companies), and outputs the filtered results in a structured CSV file.

Methodology

Tools & Technologies Used

- **Python 3.10**
- **Biopython (Entrez)** – for accessing PubMed API
- **Poetry** – dependency and package management
- **argparse** – for building CLI
- **CSV module** – for output
- **Type hints** – for static type checking

Data Flow

1. **Input Query**
   CLI accepts any PubMed query string (e.g., "gene therapy 2023").
2. **Data Fetching**
   - Utilizes Bio.Entrez to fetch metadata for relevant papers.
   - Extracts paper ID, title, publication date, authors, affiliations.
3. **Filtering Heuristic for Non-Academic Authors**
   - Affiliation strings are scanned for academic keywords:

     "university", "college", "institute", "department", "school", etc.
   - If no academic keyword is found, the author is marked **non-academic**.
   - Emails are optionally extracted if present in the affiliation text.
4. **Output**
   - Results are written to a CSV file with:
     - Pubmed ID
     - Title

- Publication Date
- Non-Academic Authors
- Company Affiliation(s)
- Corresponding Author Email

Features Implemented

Feature	Description
Search any PubMed query	Through CLI
Filter non-academic authors	Using affiliation heuristic
Export results to CSV	Filename can be customized
Optional CLI flags	--max-results, --debug, --file
Typed Python	All functions and data structures include type hints
Modular Code Structure	cli.py for interface, pubmed.py for logic
Project managed using Poetry	For reproducible builds

Example Queries & Results

Query Used	Papers Found	Non-Academic Authors
"gene therapy"	25	14 filtered and saved
"cancer drug"	20	10 filtered
"cell therapy AND Novartis"	20	13 filtered

Files were saved as:

- results_gene.csv
- results_cell.csv
- output_gene_therapy.xlsx

Testing & Validation

- Ran multiple PubMed queries and verified:
  - CSV content matches paper metadata.
  - Filtering accurately removes academic-only affiliations.

- Edge cases handled:
    - No non-academic authors → empty CSV or skipped entries.
    - Affiliation strings with mixed keywords.

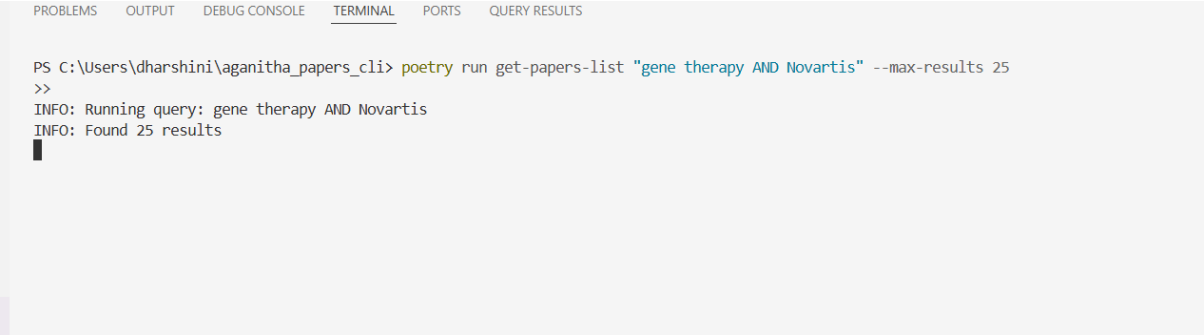
Summary

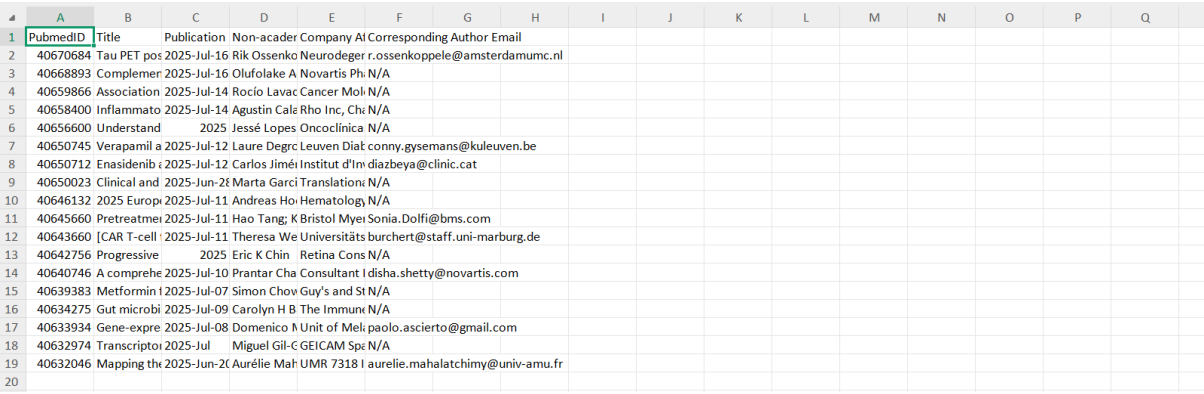
This CLI tool is cleanly structured, typed, tested, and built with maintainability in mind. It meets the core requirements of the assignment:

- Programmatic search via PubMed
- Filtering for non-academic authors
- CSV output
- Typed Python
- CLI interface with user customization

The project has been tested and uploaded in Github. A short demo video is also attached along with this document.

Output





	A	B	C	D	E	F	G	H
1	PubmedID	Title	Publication	Non-acader	Company Af	Corresponding Author Email		
2	40670684	Tau PET pos	2025-Jul-16	Rik Ossenko	Neurodeger	r.ossenkoppele@amsterdamumc.nl		
3	40668893	Complemer	2025-Jul-16	Olufolake A	Novartis Ph	N/A		
4	40659866	Association	2025-Jul-14	Rocío Lavac	Cancer Mol	N/A		
5	40658400	Inflammato	2025-Jul-14	Agustin Cale	Rho Inc, Cha	N/A		
6	40656600	Understand	2025	Jessé Lopes	Oncoclínica	N/A		
7	40650745	Verapamil a	2025-Jul-12	Laure Degrc	Leuven Diak	conny.gysemans@kuleuven.be		
8	40650712	Enasidenib (	2025-Jul-12	Carlos Jiméi	Institut d'In	diazbeya@clinic.cat		
9	40650023	Clinical and	2025-Jun-28	Marta Garci	Translationa	N/A		
10	40646132	2025 Europ	2025-Jul-11	Andreas Ho	Hematology	N/A		
11	40645660	Pretreatmer	2025-Jul-11	Hao Tang; K	Bristol Myer	Sonia.Dolfi@bms.com		
12	40643660	[CAR T-cell	2025-Jul-11	Theresa We	Universitäts	burchert@staff.uni-marburg.de		
13	40642756	Progressive	2025	Eric K Chin	Retina Cons	N/A		
14	40640746	A comprehe	2025-Jul-10	Prantar Cha	Consultant I	disha.shetty@novartis.com		
15	40639383	Metformin I	2025-Jul-07	Simon Chow	Guy's and St	N/A		
16	40634275	Gut microbi	2025-Jul-09	Carolyn H B	The Immune	N/A		
17	40633934	Gene-expre	2025-Jul-08	Domenico M	Unit of Mel	paolo.ascierto@gmail.com		
18	40632974	Transcriptor	2025-Jul	Miguel Gil-G	GEICAM Spa	N/A		
19	40632046	Mapping the	2025-Jun-2(	Aurélie Mah	UMR 7318 I	aurelie.mahalatchimy@univ-amu.fr		
20								

Github: https://github.com/codewithdharsh/aganitha-pubmed-cli

Linkedin: https://www.linkedin.com/in/dharshini-c-4a0b5825a/

Vedio: https://drive.google.com/file/d/1E-G2FKq-a5uHoGf37hhS5T3su8g4WvZ8/view?usp=drivesdk