PubMed Fetcher Project Report

Overview:

This project involves building a Python-based command-line tool that fetches research papers from PubMed using a user-defined query. The tool filters papers with at least one non-academic author affiliated with a pharmaceutical or biotech company, and outputs the result in a CSV file.

Objective:

- Accept a user-defined PubMed search query.
- Identify research papers that include authors from non-academic and pharma/biotech affiliations.
- Output relevant details into a structured CSV format.

Approach:

1. PubMed API Integration:
   - Used NCBI's E-Utilities API:
     - esearch.fcgi: To fetch paper IDs.
     - efetch.fcgi: To fetch article metadata in XML format.

2. Command-line Interface:
   - Built with argparse.
   - Supports -h, -f, -d flags.

3. Author Filtering Heuristics:
   - Academic exclusion: Matches university, college, institute.
   - Company inclusion: Matches pharma, biotech, inc, gmbh, etc.
   - Extracts corresponding author email via regex.

4. Modular Design:

- fetch.py: API calls.
- utils.py: Filtering heuristics.
- main.py: CLI handling.

Methodology:

- For a given query:

1. Search using esearch.
2. Fetch article data using efetch.
3. Parse authors and affiliations.
4. Output a CSV with relevant columns.

Output Format:

CSV Columns:

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

Tools Used:

- Python 3.9+
- Poetry
- Requests
- Regex

- Argparse

Results:

- Functional with test queries like "covid vaccine".

- Outputs well-formatted CSV with filtered authors.

Evaluation Summary:

Functional Requirements:

Typed Python:

Efficient API Calls:

CLI Features:

Modular Code:

Robust Error Handling:

Bonus - Modular Split:

Bonus - Test PyPI Ready: Pending

Future Work:

- Add pagination, caching.

- Publish to PyPI.

- Unit tests.

- Notifications on completion.

Author: Your Name

Email: your_email@example.com