**PubMed Paper Fetcher - Project Report**

**Overview:**

This project implements a command-line Python tool to fetch research papers from PubMed using a user-provided query and extract papers that have at least one non-academic author, such as someone affiliated with a pharmaceutical or biotech company.

**Approach:**

1. Command-Line Interface (CLI):

- Built using argparse to accept a search query.

- Supports optional flags: --debug for logs and --file to export results.

2. PubMed API Integration:

- Used Entrez from biopython to search (esearch) and fetch (efetch) paper metadata.

- A valid user email is configured to comply with API access.

3. Filtering Logic:

- Parsed author affiliations to detect non-academic affiliations using a keyword-based heuristic.

- Keywords like university, institute, school, hospital, and college were used to classify academic affiliations.

- Affiliations not containing these were considered non-academic, likely indicating company involvement.

4. Data Extraction:

- From each paper, the following fields were extracted:

- PubMed ID

- Title

- Publication Year

- Non-academic Author(s)

- Company Affiliation(s)

- Corresponding Author Email (if available)

5. Export:

- Results were either printed to console or saved as a .csv file using pandas.

**Methodology:**

- Tech Stack:

1. Python 3.12
2. Poetry for dependency management and packaging
3. Biopython for PubMed API
4. Pandas for CSV handling

**Code Structure:**

1. fetch.py : PubMed data fetching
2. filter.py : Filtering logic for non-academic authors
3. csv\_writer.py : Handles CSV export
4. cli.py :Main entry point for command-line interaction

**Typing and Documentation:**

- Code is written with type hints and includes meaningful variable names and comments for readability and maintainability.

**Results:**

- Successfully fetched and filtered PubMed research papers based on real queries like 'cancer immunotherapy' and 'diabetes treatment'.

- Identified and exported papers with company-affiliated authors, producing structured CSV reports.

- Detected companies such as 'Genentech, Inc.', 'Pfizer Global Research and Development', etc.

- Tool handled missing or malformed data gracefully.

**Repository:**

GitHub: https://github.com/baranibh/pubmed-fetcher