# Research Paper Fetcher – Project Report

## 1. Introduction

Research in the pharmaceutical and biotech industry relies heavily on academic and industry publications. This project aims to develop a Python-based command-line tool that fetches research papers from PubMed, filtering results to include those affiliated with pharmaceutical or biotech companies. The tool processes and exports data in a structured CSV format.

## 2. Objectives

The main objectives of this project are:

* Fetch research papers based on user queries from PubMed
* Identify papers with at least one non-academic author
* Extract relevant details like PubMed ID, Title, Publication Date, and Author Information
* Provide a command-line interface for easy interaction
* Export results to a CSV file

## 3. Methodology

### 3.1 Tools and Technologies Used

- Programming Language: Python (≥3.10)

- API Used: PubMed Entrez API

- Package Management: Poetry

- Libraries:

* requests – For making API requests
* pandas – For data processing
* argparse – For handling command-line arguments

### 3.2 System Architecture

The project follows a modular design with two key components:

* -Core Module (pubmed\_fetcher.py): Handles API requests and data extraction
* - CLI Module (cli.py): Provides a command-line interface for executing queries

### 3.3 Implementation Details

1. Fetching Data from PubMed

- The tool sends a GET request to PubMed’s Entrez API.

- Retrieves paper metadata such as PubMed ID, Title, and Authors.

2. Filtering Non-Academic Authors

- Authors are checked for affiliations with companies instead of universities.

- Heuristics such as email domains (@company.com) or keywords (Pharma, Biotech) are used.

3. Saving Data to CSV

- Extracted results are formatted into a Pandas DataFrame.

- The data is exported as a CSV file if specified.

## 4. Results and Output

### 4.1 Example Query Execution

Command:

poetry run get-papers-list "cancer treatment" -f results.csv

### 4.2 Example Output

Console Output:

[  
 {"PubmedID": "12345678", "Title": "Cancer Treatment Advancements", "Publication Date": "2024-02-01", "Company Affiliation": "Pfizer"},  
 {"PubmedID": "87654321", "Title": "New Drug Research", "Publication Date": "2024-01-15", "Company Affiliation": "Moderna"}  
]

CSV File Output (results.csv):

PubmedID,Title,Publication Date,Company Affiliation  
12345678,Cancer Treatment Advancements,2024-02-01,Pfizer  
87654321,New Drug Research,2024-01-15,Moderna

## 5. Challenges and Solutions

|  |  |
| --- | --- |
| Challenges | Solutions Implemented |
| API Rate Limits | Implemented pagination and error handling for API requests |
| Identifying Non-Academic Authors | Used heuristic filtering (checking email domains & company names) |
| Large Data Processing | Optimized API requests and used Pandas for efficient processing |

## 6. Publishing to TestPyPI

This package is published on TestPyPI. To publish new versions:

1. Build the package:

poetry build

2. Upload to TestPyPI:

twine upload --repository testpypi dist/\*

3. Verify the upload:

Visit https://test.pypi.org/project/pubmed-paper-fetcher-abhishek/ to confirm.

## 7. Conclusion and Future Work

This project successfully implements a Python-based CLI tool for fetching and filtering research papers from PubMed.

🔹 Key Achievements:

* Automated research paper retrieval
* Filtered company-affiliated authors
* Provided a CLI interface for easy use
* Exported results in CSV format

🔹 Future Enhancements:

* Expand heuristic methods for non-academic authors
* Implement a GUI-based version for better accessibility
* Add advanced filtering options (e.g., keyword-based topic extraction)

## 8. References

• PubMed API Documentation: https://www.ncbi.nlm.nih.gov/home/develop/api/

• Python Poetry: https://python-poetry.org/docs/

• Pandas Library: https://pandas.pydata.org/