**📄 Project Report: PubMed Non-Academic Scraper**

**1. Project Title**

**PubMed Non-Academic Scraper CLI Tool**

**2. Objective**

The objective of this project is to develop a Command-Line Interface (CLI) tool that fetches research articles from PubMed based on a user-defined query and filters them to identify papers with **non-academic affiliations**. This tool helps highlight industry participation in specific research domains.

**3. Tools & Technologies Used**

* **Language**: Python 3.10+
* **Libraries**:
  + requests – for API communication
  + beautifulsoup4 – for HTML/XML parsing
  + typer – for building CLI interface
  + pytest – for testing
* **Packaging & Publishing**:
  + poetry – for dependency and packaging management
  + Published to [Test PyPI](https://test.pypi.org/project/my-first-pubmed-project/)
* **Version Control**: Git & GitHub

**4. Methodology**

**a. CLI Interface**

The CLI is built using Typer, allowing users to enter a query string, specify an optional output file, and enable debugging.

**b. PubMed API Integration**

The tool uses the NCBI Entrez E-utilities API (esearch and efetch) to:

* Retrieve IDs based on the search query
* Fetch full article metadata (authors, affiliations, etc.)

**c. Filtering Logic**

* A custom logic is implemented to **filter authors** whose affiliations contain **non-academic keywords** like:
  + "Inc", "LLC", "Technologies", "Solutions", "Company", "Ltd", etc.
* This identifies individuals from the **corporate or industry** sector rather than universities or academic institutions.

**d. Data Export**

* The filtered results are saved as a structured CSV file including:
  + Title
  + Authors
  + Affiliations
  + DOI / PubMed Link
  + Non-academic author names
  + Emails (if available)

**5. Example Use Case**

get-papers-list "AI in Healthcare" --output ai\_healthcare.csv

This command will output a CSV file containing industry-related research contributions in the "AI in Healthcare" domain.

**6. Results**

* The tool successfully identified **non-academic contributors** in various PubMed queries.
* Exported data was clean and ready for downstream processing or visualization.
* Demonstrated the feasibility of **analyzing industry-academia collaboration** using publicly available metadata.

**7. Outcome & Significance**

This CLI tool provides a fast, automated method to study non-academic participation in scientific literature, especially useful for:

* Policy analysts
* Industry researchers
* Academics exploring collaboration patterns

**8. Future Work**

* Add GUI interface for easier access
* Extend support to full-text scraping (if permitted)
* Train an ML model to auto-classify affiliations with higher accuracy

**9. Author**

**Srishanth M**