PubMed Research Paper Fetching Tool - Summary Report

1. Introduction

This report provides an overview of the approach, methodology, and results of the **PubMed Research Paper Fetching Tool**, which retrieves research papers from the **PubMed API** and filters results based on non-academic author affiliations.

2. Approach

The tool is implemented as a **Python CLI application** that interacts with the **PubMed API** to retrieve research papers based on a user-specified query. The extracted data is stored in a structured CSV format for easy access and analysis.

3. Methodology

3.1. Data Retrieval

- The tool queries **PubMed** using esearch.fcgi to fetch relevant **paper IDs**.
- It then retrieves metadata for these papers using esummary.fcgi.
- For deeper metadata extraction (author affiliations and emails), efetch.fcgi is used to obtain **full-text XML** data.

3.2. Filtering Non-Academic Authors

- The script checks author affiliations for **keywords** related to **companies** (e.g., "Pharma", "Biotech", "Inc", etc.).
- If an author's affiliation matches company-related terms, they are classified as **non-academic authors**.

3.3. Extracting Corresponding Author Email

- Emails are extracted from **structured XML fields** using the efetch.fcgi response.
- The script scans **Author > ContactInfo > Email** fields to retrieve the corresponding author's email.

3.4. Data Storage

- Extracted data is stored in a **CSV file** with the following columns:
    - **PubmedID**

    - **Title**
    - **Publication Date**
    - **Non-academic Author(s)**
    - **Company Affiliation(s)**
    - **Corresponding Author Email**
- A command-line option allows users to specify the output filename (-f output.csv).

4. Results

- The tool successfully retrieves and filters research papers.
- **Limitations:** Some records return "N/A" for emails or affiliations due to missing metadata in PubMed.
- **Enhancements made:** Improved XML parsing for better email and affiliation detection.

5. Conclusion

This tool provides an automated way to fetch and filter **PubMed** research papers based on non-academic author affiliations. Future improvements may include **more robust company name matching** and **handling missing email data with alternative sources**.

Developed by: DEGA NIKHITHA
Date: 21-03-2025