**PubMed Research Paper Fetcher**

**Overview**

The **PubMed Research Paper Fetcher** is a Python-based tool designed to retrieve research papers from PubMed using the Entrez API. This tool allows filtering papers based on author affiliation, specifically identifying non-academic authors using heuristics such as email domains and company affiliations. The filtered results are then saved in a structured CSV file.

**Objectives**

* **Fetch** research papers related to a given query from PubMed.
* **Filter** papers to retain those with non-academic authors.
* **Extract** useful metadata like corresponding author emails and affiliations.
* **Save** the filtered results in a CSV file for further analysis.

**Features**

* Uses **Entrez API** (esearch and efetch) to query PubMed.
* Filters **non-academic authors** based on company affiliation and email heuristics.
* Supports **CSV export** for structured data storage.
* Implements **error handling** for missing or incomplete metadata.

**Project Structure**

PubMed-Fetcher/

├── fetch\_papers.py # Core script for fetching papers

├── filter\_papers.py # Script for filtering non-academic authors

├── save\_to\_csv.py # Saves filtered results into a CSV file

├── get\_papers\_list.py # Main execution script

├── requirements.txt # Dependencies

├── README.md # Documentation

**How It Works**

1. **Search Query Execution:**
   * The script takes a query (e.g., "cancer research") and fetches paper IDs from PubMed using the esearch API.
   * Example API call:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=cancer+research&retmax=10&retmode=json

1. **Fetching Full Article Metadata:**
   * Using the article IDs from esearch, the script retrieves detailed metadata with efetch.
   * Example API call:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=<ID1>,<ID2>&retmode=xml

1. **Filtering Non-Academic Authors:**
   * Heuristics applied:
     + If an author's email **does not** contain .edu, .ac, .gov, or .org, it is considered non-academic.
     + If the affiliation contains "Inc.", "Ltd.", "Corp.", etc., it is considered a company.
2. **Saving Results to CSV:**
   * Extracted metadata (ID, title, authors, affiliation, email) is stored in filtered\_papers.csv.

**Installation & Usage**

**Prerequisites**

* Python 3.9+
* Required libraries: requests, csv
* API access to **NCBI Entrez**

**Installation**

pip install -r requirements.txt

**Running the Script**

python get\_papers\_list.py --query "cancer research" --file output.csv

**Challenges & Improvements**

**Current Challenges**

* Some papers **do not provide** author email or affiliation.
* The API **rate limits requests**, requiring careful execution.

**Future Improvements**

* Implement **better NLP techniques** to classify affiliations.
* Support **additional data sources** beyond PubMed.
* Enhance UI with a **web-based search interface**.

**Conclusion**

This project is an automated solution for fetching and filtering research papers from PubMed. By improving the filtering process and data extraction techniques, it can be a valuable tool for identifying industry contributions to scientific research.