**Report on PubMed Data Extraction Project**

**Prepared by: Singuluri Srilatha**

**Date:9/03/2025**

**1. Introduction**

This report outlines the development, execution, and validation of the PubMed Data Extraction Script, which retrieves research papers based on specified search terms. The script is designed to filter out papers with purely academic affiliations and retain those with non-academic authors and company affiliations, exporting the results to a structured CSV file.

**2. Project Objectives**

The primary objectives of this project were:

* To extract PubMed research papers based on search queries.
* To filter results to include only papers with non-academic authors or corporate affiliations.
* To generate a structured CSV file containing:
  + PubmedID (Unique Identifier)
  + Title (Paper Title)
  + Publication Date
  + Non-Academic Author(s)
  + Company Affiliation(s)
  + Corresponding Author Email

**3.Implementation Details**

**3.1 Technologies Used**

* **Programming Language: Python**
* **Libraries Used:** 
  + **Bio.Entrez (for PubMed API requests)**
  + **csv (for exporting results)**
  + **re (for filtering author affiliations)**
  + **pandas (for data processing)**

**3.2 Functionality Overview**

The script follows these steps:

1. User Input: Accepts a search term (e.g., *"cancer treatment"*) as a parameter.
2. Data Retrieval: Queries the PubMed API and fetches research articles.
3. Filtering Logic:
   * Extracts author affiliations.
   * Filters out purely academic papers.
   * Retains only papers with non-academic authors or company affiliations.
4. Data Storage: Saves the filtered data as a CSV file.

**4. Challenges & Solutions**

**4.1 File Permission Errors**

* **Issue:** A PermissionError: [Errno 13] was encountered when writing to pubmed\_results.csv.
* **Solution:** Ensured the file was not in use, removed read-only restrictions, and rerouted output to a different directory.

**4.2 Filtering Accuracy**

* **Issue:** Some research papers had ambiguous author affiliations.
* **Solution:** Implemented **regular expressions** to detect **company names** in author details, refining the selection criteria.

**4.3 API Rate Limits**

* **Issue:** Excessive API calls resulted in PubMed rate-limiting errors.
* **Solution:** Introduced **delays between API requests** to comply with PubMed’s guidelines

**5. Validation & Testing**

* The script was tested on multiple search terms. The output was reviewed to ensure that:  
  ✔️ **Only papers with non-academic authors or company affiliations were included**.  
  ✔️ **Academic-only papers were excluded** as intended.  
  ✔️ **The exported CSV contained all required fields** in the correct format.

**6. Conclusion**

This project successfully implemented a **PubMed data extraction system** with filtering capabilities. The script meets the specified requirements, ensuring that only relevant research articles with **industry affiliations** are extracted.

The **final implementation** is functional and ready for deployment. The script can be further improved by:

* Expanding company name databases for better affiliation detection.
* Enhancing error handling for large-scale queries.
* Integrating visualization tools for better insights.