PubMed Research Paper Fetcher: Report

Introduction

The PubMed Research Paper Fetcher is a Python-based tool designed to retrieve research papers from PubMed using their unique identifiers (PubMed IDs). It extracts key details such as titles, authors, publication dates, affiliations, and identifies industry-affiliated authors. The extracted data is then saved in CSV format for further analysis.

Approach

a. Fetching Data from PubMed

- The tool uses the **NCBI E-utilities API** to fetch research papers.
- The `requests` library is used to send HTTP GET requests to retrieve paper metadata in XML format.

b. Parsing and Extracting Information

- The XML response is parsed using `xml.etree.ElementTree`.
- Key elements extracted include:
  - **Title:** Extracted from the `<ArticleTitle>` tag.
  - **Publication Date:** Retrieved from the `<PubDate>` field.
  - **Authors and Affiliations:** Extracted from the `<AuthorList>` tag, where each author's last name, fore name, and affiliation are identified.
  - **Non-Academic Authors:** Identified by checking affiliations for keywords like "Pharma", "Biotech", "Inc", etc.
  - **Corresponding Author Email:** Retrieved if available.

c. Saving Results

- Extracted data is structured into a Python dictionary.
- Data is saved as a CSV file using the `csv` module for easy analysis.

3. Methodology

a. Input Method

- The script accepts a **PubMed ID** as input via the command line.
- Example usage:

```
python fetch_pubmed_papers.py  39725271
```

- Alternative input method using `curl`:

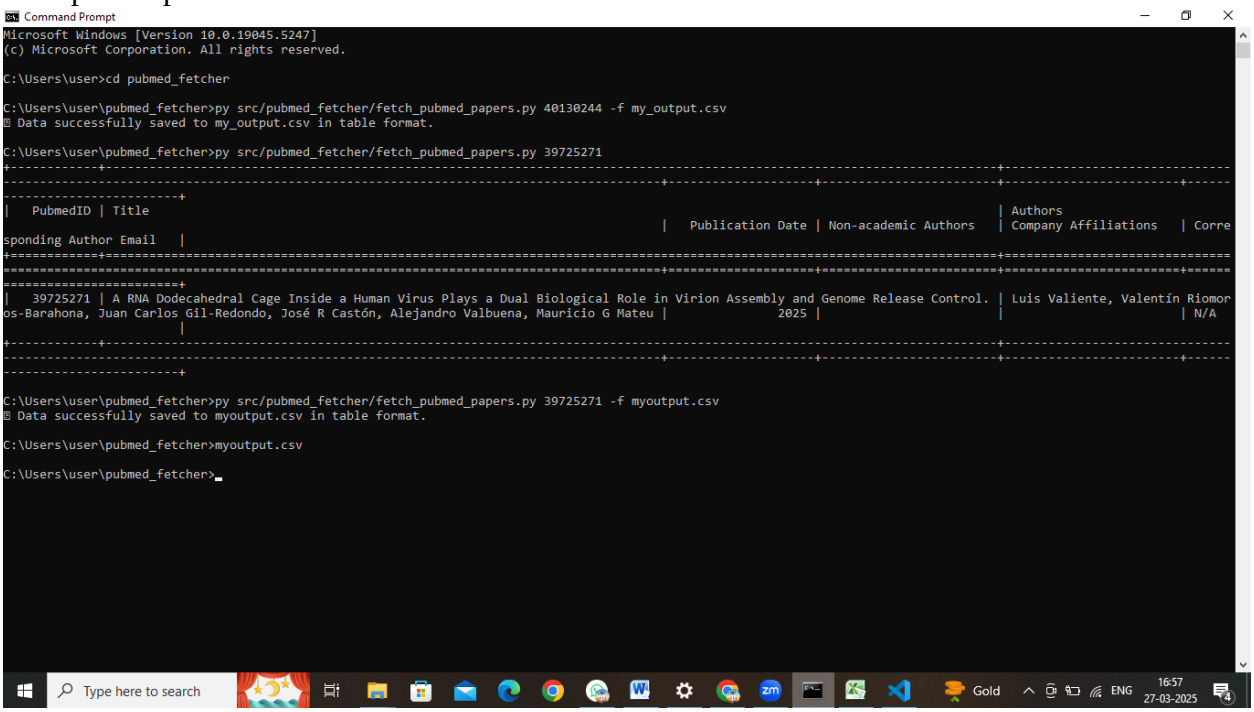
```
curl -k http://localhost:5000/fetch?pubmed_id=40130244
```

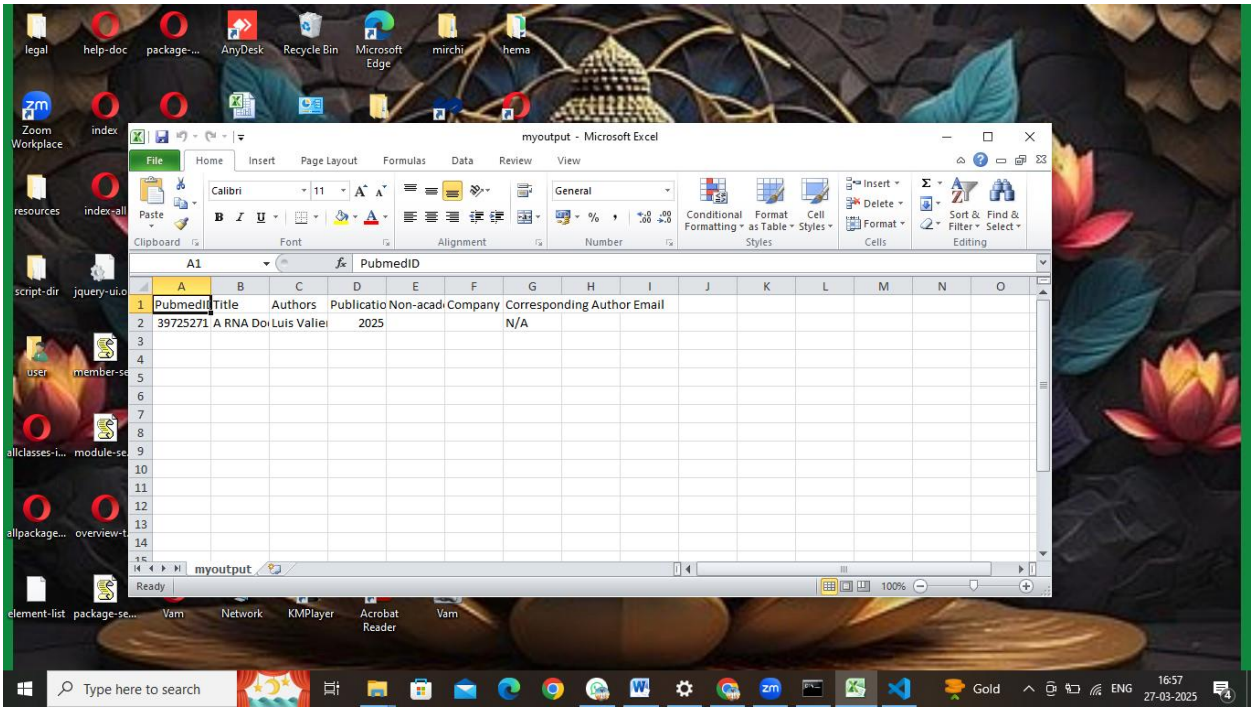
b. Processing Steps

1. Validate and process the input.
2. Fetch XML data from PubMed.
3. Parse XML to extract key details.
4. Identify industry-affiliated authors.
5. Store results in CSV format.

c. Output Format

- CSV file with the following columns:
  - **PubMed ID**
  - **Title**
  - **Authors**
  - **Publication Date**
  - **Non-Academic Authors**
  - **Company Affiliations**
  - **Corresponding Author Email**
- Example output:
-



- 

4. Results

- Successfully retrieved and processed multiple PubMed research papers.
- Identified industry-affiliated authors based on specified keywords.
- Exported structured CSV reports for further analysis.

5. Conclusion

The PubMed Research Paper Fetcher effectively automates the retrieval and filtering of research papers. This tool is useful for researchers and analysts looking to track industry involvement in academic research.

6. Future Improvements

- Support for batch processing of multiple PubMed IDs.
- Integration with a database for historical tracking.
- Improved filtering based on AI-based entity recognition.
- Web-based interface for easier accessibility.