PubMed Paper Filtering Report

Project Title: Identifying and Filtering Non-Academic Papers from PubMed

Date: 10th March 2025

Author: MADHUMITA PRADHAN

Introduction

The goal of this project is to filter non-academic papers from PubMed search results using Python. PubMed is a vast repository of scientific literature, and distinguishing between academic and non-academic papers is crucial for research accuracy. This report summarizes the approach, methodology, and results obtained from implementing a filtering system.

Approach

The project is structured into several steps:

1. **Fetching papers from PubMed API**: Using requests to retrieve metadata for a given query.
2. **Extracting relevant details**: Parsing XML responses to get paper titles, authors, publication dates, and affiliations.
3. **Filtering non-academic papers**: Identifying papers with affiliations related to corporations, biotech companies, and other non-academic institutions.
4. **Saving results**: Writing filtered results into a structured CSV file for easy analysis

---

Methodology

3.1 Data Collection

- The PubMed API (eutils) was used to search for papers using a given query (e.g., "cancer drug development").
- The esearch endpoint retrieved a list of PubMed IDs.
- The efetch endpoint was used to obtain detailed metadata for each paper.

3.2 Data Processing

- The XML response was parsed to extract:
  - **Title**
  - **Publication Date**
  - **Authors**
  - **Author Affiliations**
- A function was implemented to identify non-academic papers based on specific keywords in affiliations (e.g., "Inc.", "Ltd.", "Pharmaceutical", "Tech").

3.3 Filtering Criteria

- Papers were classified as "non-academic" if at least one author had an affiliation that matched corporate keywords.
- The script checked for affiliations using a predefined list of non-academic keywords.

3.4 Storing Results

- Filtered results were saved to results.csv with the following columns:
  - **PubMed ID**
  - **Title**
  - **Publication Date**
  - **Non-Academic Authors**
  - **Company Affiliation**

Results

The script successfully filtered non-academic papers based on author affiliations. Below is a sample of the output:

PubMed ID	Title	Publication Date	Non-Academic Authors	Company Affiliation
40059423	Progress and Application of Multifunctional Ultrasound Theranostic Agents.	07-Mar-2025	Fang; Lei	Chongqing Engineering and Technology Research Center
40059421	Advancements and Challenges of Plant-derived Extracellular Vesicles in Anti-Cancer Strategies and Drug Delivery.	07-Mar-2025	Liu	Shenzhen Shifangjie Technology Co., Ltd.

Conclusion

The implemented pipeline effectively identified and filtered non-academic papers from PubMed search results. The methodology provides a scalable approach to analyzing scientific literature and distinguishing between academic and industry-driven research. Further improvements could include refining the keyword list and incorporating machine learning for better classification.

Future Improvements

- Expand the keyword list to improve accuracy.
- Implement a confidence score for classification.
- Develop a web-based interface for user-friendly interaction.

End of Report