PubMed Big Data Analysis: Project Report

Mikhail Sumskoi Constantinos Odysseos

May 2025

1 Introduction

The rapid growth of biomedical literature presents significant challenges for researchers seeking to retrieve and synthesize relevant information. The PubMed database, maintained by the National Library of Medicine, hosts over 36 million articles, making manual search and analysis impractical. This project addresses these challenges by leveraging Big Data technologies, Natural Language Processing (NLP), and graph analytics to enable intelligent information retrieval and knowledge discovery. By automating data retrieval, preprocessing, and advanced analysis, the project aims to support biomedical professionals in uncovering semantic insights, influence metrics, and research trends. The pipeline, built on Apache Spark, GraphFrames, and modern NLP tools, ensures scalability and reproducibility, offering a robust foundation for academic and practical applications.

2 Objectives

The project focuses on four primary objectives to enhance biomedical research discovery:

- **Automated Data Retrieval and Preprocessing**: Develop a scalable pipeline to extract and clean PubMed data, converting XML files into Parquet format for efficient processing with Apache Spark. This ensures rapid access to metadata, abstracts, and citation information.
- **Semantic and Influence Analysis**: Extract semantic topics using Latent Dirichlet Allocation (LDA) and compute influence metrics, such as citation counts and journal impact factors, to identify high-impact research.
- **Graph-Based Representations**: Construct citation and authorship networks using GraphFrames to model relationships between articles and researchers, enabling the identification of influential contributors and research clusters.
- **Efficient Research Discovery**: Provide biomedical professionals with tools for high-quality, automated research discovery, including topic modeling, gene ontology tagging, and high-impact article filtering.

3 Data Processing and Exploratory Data Analysis

The source data comprises PubMed XML files, which were converted to Parquet format to optimize storage and querying with Apache Spark. The core dataset includes approximately 25,000 articles, filtered to ensure the presence of abstracts and complete metadata (e.g., PMID, title, journal, and publication date). The conversion process, implemented in the `SCHEMA_SELECTION.ipynb` notebook, involved parsing nested XML structures and defining a consistent schema for downstream analysis.

Exploratory Data Analysis (EDA), conducted in the `EDA.ipynb` notebook, provided insights into the datasets characteristics. Key tasks included:

- **Parsing Nested XML**: Extracted metadata fields, such as authors, keywords, and publication dates, resolving inconsistencies in XML structure.
- **Handling Missing Values**: Imputed or removed records with missing abstracts or critical metadata, ensuring data quality.
- **Citation Frequency Analysis**: Analyzed citation distributions to identify highly cited articles, revealing a skewed distribution where a small fraction of articles accounted for most citations.
- `Term Distributions`: Examined word frequencies in abstracts, identifying common biomedical terms like “cancer, “gene, and “protein.
- `Publication Patterns`: Visualized trends in publication volume over time, noting a steady increase in articles published annually.

These findings guided subsequent analyses by highlighting data quality issues and research trends.

4 Advanced Analysis

The project employed several advanced techniques to extract insights from the PubMed dataset, as detailed below:

4.1 Citation Analysis

Citation data was retrieved using the PubMed E-utilities API and stored in `pmid_citations_901.csv`. The `Automated_Citation_Retrieval_with_PubMed_E_utilities_API_Final_Version` notebook automated this process, enriching article metadata with citation counts and references. This enabled the computation of influence metrics, such as the number of times an article was cited, which served as a proxy for its impact within the biomedical community.

4.2 GraphFrames Analysis

Citation and authorship networks were constructed using GraphFrames in the `PUBMED_GRAPHFRAMES.ipy` and `PUBMED_Spark_GraphFrames.ipynb` notebooks. The citation network, a directed graph with over 12,000 edges, modeled articles as nodes and citations as edges. PageRank and centrality measures identified influential articles and journals. Similarly, the author collaboration graph connected researchers based on co-authorship, revealing key contributors and research clusters. For example, highly central authors were often associated with multiple high-impact publications.

4.3 Gene Ontology Tagging

Abstracts were mapped to Gene Ontology (GO) categoriesBiological Process (BP), Molecular Function (MF), and Cellular Component (CC)using the `PUBMED_GENEONTOLOGY.ipynb` notebook. A dictionary-based approach matched biomedical terms to GO identifiers, with 72% of articles successfully tagged with at least one GO term. Common categories included “cell signaling (BP), “protein binding (MF), and “nucleus (CC), reflecting prevalent research themes.

4.4 Topic Modeling

Latent Dirichlet Allocation (LDA), implemented in the `LDA.ipynb` notebook, extracted 10 dominant biomedical themes from article abstracts. These included:

- Cancer research (e.g., terms like “tumor, “oncogene)
- Inflammation and immune response
- Neural pathways and neurodegenerative diseases
- Gene regulation and epigenetics
- Cardiovascular health
- Infectious diseases
- Metabolic disorders
- Drug development
- Protein interactions
- Stem cell research

Each topic was represented by a distribution of terms, enabling researchers to explore thematic trends and identify relevant articles.

4.5 High-Impact Filtering

A dataset of 539 articles published in journals with an impact factor of 20 or higher was curated and stored in `high_impact_pmid_journal.parquet`. The list of high-impact journals, sourced from `impact factor journals.xlsx`, included titles like *Nature*, *Science*, and *The Lancet*. This filtering enabled targeted analysis of top-tier publications.

5 Summary of Key Results

The project yielded several significant outcomes:

- A directed citation network with over 12,000 edges, enabling the identification of influential articles and authors via PageRank and centrality.
- Ten dominant biomedical research themes extracted via LDA, providing a thematic overview of the dataset.
- 72% of abstracts mapped to Gene Ontology categories, facilitating semantic categorization.

- 539 articles identified from high-impact journals, supporting focused analysis of top-tier research.

These results demonstrate the pipelines ability to uncover actionable insights from large-scale biomedical data.

6 Project Files and Structure

The project repository (`https://github.com/Cody9494/PubMed_BigData_Analysis`) is organized as follows:

- `PubMedData/parquet/`: Stores cleaned PubMed data in Parquet format for efficient querying.
- `SCHEMA_SELECTION.ipynb`: Defines the schema for XML-to-Parquet conversion and metadata structure.
- `Automated_Citation_Retrieval_with_PubMed_E_utilities_API_Final_Versio` Automates citation data retrieval using the PubMed API.
- `EDA.ipynb`: Performs exploratory data analysis, including citation and term distribution analyses.
- `PUBMED_GENEONTOLOGY.ipynb`: Maps abstract terms to Gene Ontology categories.
- `LDA.ipynb`: Conducts topic modeling on article abstracts.
- `PUBMED_GRAPHFRAMES.ipynb` and `PUBMED_Spark_GraphFrames.ipynb`: Build and analyze citation and author networks.
- `high_impact_pmid_journal.parquet`: Contains a pre-filtered list of high-impact publications.
- `impact factor journals.xlsx`: Lists high-impact journals with impact factors.
- `pmid_citations_901.csv`: Stores citation data by PMID.

This modular structure ensures reproducibility and ease of use.

7 Conclusion

This project delivers a scalable, reproducible pipeline for analyzing PubMed data, integrating automated retrieval, semantic modeling, and graph-based analytics. By leveraging Apache Spark, GraphFrames, and NLP techniques, it addresses the challenges of information overload in biomedical research. The pipeline supports advanced discovery through topic modeling, gene ontology tagging, and influence analysis, enabling researchers to identify high-impact articles, uncover thematic trends, and explore research networks. Future work could extend the pipeline to include real-time data updates, incorporate additional ontologies, or integrate machine learning for predictive analytics, further enhancing its utility for biomedical professionals.