MIT805-Assignment1

u23825848

August 2023

1 Data decision, collection and processes

Dataset: PubMed Knowledge Graph
Size: 11 Gig
Format: The files are CSV; Information is alpha numeric.
Age: Last updated 2 years ago (07 January 2022 at 20:20:27)

1.1 Overview of the PubMed Knowledge Graph

Introduction
The PubMed Knowledge Graph, which can be accessed through the PubMed database, is a thorough and dynamic depiction of biomedical data that has been drawn from the enormous body of scientific literature. It makes use of cutting-edge graph database technology and natural language processing techniques to organize, link, and make sense of the various biomedical entities and relationships discovered in research publications.

Key Elements and Characteristics:
1. Data processing and ingestion: The millions of scientific papers saved in the PubMed database's textual content are processed to create the knowledge graph. These articles cover a broad range of biomedical subjects. Techniques for natural language processing are employed to separate the structured data from the unstructured text.

2. Entity Extraction and Linking: The knowledge graph recognizes, and links particular entities described in the articles to standardized concepts through named entity identification and disambiguation. To ensure correct representation and information retrieval, this method entails linking synonyms, aliases, and related terms. Metadata enhanced with IDs and semantic type information is added to the entities.

3. Relationship Establishment: A sophisticated network of linkages that resembles real-world biomedical associations connects the retrieved components. The knowledge graph offers insights into the interconnectedness of biomedical concepts and events by capturing these relationships.

4. Graph database architecture: The PubMed Knowledge Graph is designed as a graph database, enabling effective interconnected data archiving, retrieval, and querying. Given the numerous linkages between elements in biomedical information, graph databases are particularly adept at modelling complex relationships. With the help of this design, researchers may more quickly study linkages and patterns in the graph.

5. Semantic Enrichment: Semantic enrichment techniques are used to increase the usefulness of the knowledge graph. To provide standardized and consistent language, this entails incorporating ontologies and controlled vocabularies. This improves the graph's interoperability and usefulness by ensuring that it can be queried and understood using standard biomedical concepts.

6. Use in Biomedical Research: The PubMed Knowledge Graph is a valuable resource for researchers, doctors, and other biomedical specialists. By allowing users to navigate the graph to locate pertinent relationships between genes, diseases, treatments, and more, it facilitates literature mining, hypothesis formulation, and knowledge discovery. It can help in finding new associations, supporting established ideas, and directing experimental plans.

7. Constant Updating and Expansion: The knowledge graph isn't a static object; it changes when new studies are published. The graph is kept up to date and represents the most recent biomedical developments thanks to frequent updates. The graph becomes a more important tool for academics looking to navigate the ever-expanding environment of biomedical knowledge as the body of scientific publications increases.

In summary, the PubMed Knowledge Graph is a revolutionary method for arranging and comprehending the vast amount of biomedical data that can be found in scientific literature. It enables researchers to find hidden correlations, speed up discovery, and enhance biomedical science by utilizing cutting-edge NLP approaches and graph database technologies. The knowledge graph is a living example of the potential for AI-driven technologies to revolutionize how we discover and utilize biomedical knowledge as it continues to develop and grow [1].

1.2 PubMed Knowledge Graph dataset demonstrates the four V's—variety, veracity, volume, and velocity—in the context of data collection and processing [2, 3].

- Variety: The PubMed Knowledge Graph dataset showcases an exceptional variety of biomedical information. The dataset spans a wide spectrum of research articles, covering topics from genetics and molecular biology to clinical medicine and epidemiology. This variety reflects the richness and complexity of the biomedical domain, enabling comprehensive insights into the interconnectedness of various biomedical concepts.
- Veracity: Ensuring the veracity of the PubMed Knowledge Graph dataset is paramount to its credibility and usability. The dataset undergoes rigorous data extraction and validation

processes, employing advanced natural language processing techniques to accurately identify and link entities from the unstructured text of research articles. Disambiguation and normalization steps ensure that entities are correctly associated with standardized concepts. In addition, semantic enrichment through ontologies and controlled vocabularies helps maintain consistent and reliable terminology across the graph. This commitment to data accuracy and reliability enhances the trustworthiness of the knowledge graph as a reliable biomedical resource.
- Volume: The PubMed Knowledge Graph dataset boasts an immense volume of data, representing a substantial portion of the scientific literature stored in the PubMed database. With millions of articles covering a wide array of biomedical topics, the dataset captures a vast and ever-expanding body of knowledge. The inclusion of a multitude of entities and relationships between them contributes to the dataset's substantial size. As new research articles are continuously published, the volume of the dataset increases, underscoring the dynamic nature of the biomedical field.
- Velocity: The velocity of the PubMed Knowledge Graph dataset pertains to the speed at which data is collected, processed, and integrated. With the ongoing publication of scientific articles, the dataset's velocity is notably high. Advanced natural language processing algorithms streamline the extraction and transformation of information from articles, enabling the rapid incorporation of new knowledge into the graph. Regular updates ensure that the dataset remains current, reflecting the latest advancements in biomedical research. This high velocity ensures that researchers have access to up-to-date information and insights.

In addition to the four V's (Variety, Veracity, Volume, and Velocity), there are a few more V's that you can use to describe the PubMed Knowledge Graph dataset: [2, 3]
- Validity: The Validity of the dataset refers to its accuracy, correctness, and fitness for the intended purpose. The PubMed Knowledge Graph dataset goes through a thorough validation process to ensure that the extracted entities, relationships, and attributes accurately represent the biomedical concepts found in the scientific literature. By adhering to established standards, controlled vocabularies, and ontologies, the dataset maintains a high level of validity, making it a reliable source of information for researchers, clinicians, and professionals in the biomedical field.
- Variability: Variability reflects the extent to which the dataset exhibits changes and fluctuations over time or across different contexts. The PubMed Knowledge Graph dataset showcases variability in terms of the evolving nature of biomedical research. New articles, discoveries, and advancements lead to changes in the dataset's content as it expands to accommodate the ever-changing landscape of scientific knowledge. This variability highlights the dataset's adaptability and relevance to ongoing research endeavours.
- Volatility: Volatility refers to the degree of instability or the rate of change within a dataset. The PubMed Knowledge Graph dataset demonstrates a certain level of volatility due to the continuous influx of new research articles. As the field of biomedicine evolves, the dataset experiences fluctuations in its content, necessitating frequent updates to maintain accuracy and relevancy. This volatility underscores the dynamic nature of the dataset and its responsiveness to emerging biomedical insights.
- Value: The Value of the dataset signifies the usefulness and significance it brings to its users. The PubMed Knowledge Graph dataset holds immense value for researchers, clini-

cians, and professionals in the biomedical domain. It serves as a powerful tool for knowledge discovery, hypothesis generation, and decision-making. By enabling the exploration of complex relationships and hidden patterns within biomedical data, the dataset adds value by facilitating informed research, accelerating discoveries, and contributing to advancements in medical science.
- Vocabulary: Vocabulary refers to the set of terms and concepts used within a dataset. The PubMed Knowledge Graph dataset employs a comprehensive biomedical vocabulary composed of standardized terms, synonyms, aliases, and controlled terminologies. This well-defined vocabulary ensures consistency and clarity in representing entities and relationships. By utilizing established ontologies and controlled vocabularies, the dataset enhances interoperability and aids in accurate data interpretation.
- Visualization: Visualization pertains to the graphical representation of data to facilitate understanding and insights. The PubMed Knowledge Graph dataset can be visualized as a complex network of interconnected entities and relationships. Visualization tools and techniques enable users to explore and comprehend the intricate web of biomedical concepts, providing a more intuitive way to navigate and analyse the wealth of information contained within the dataset.

Incorporating these additional V's—Validity, Variability, Volatility, Value, Vocabulary, and Visualization—further enriches the description of the PubMed Knowledge Graph dataset, offering a comprehensive perspective on its characteristics, significance, and impact within the biomedical research landscape.

References

(1) Xu, J.; Kim, S.; Song, M.; Jeong, M.; Kim, D.; Kang, J.; Rousseau, J. F.; Li, X.; Xu, W.; Torvik, V. I., et al. Building a PubMed knowledge graph. *Scientific data* **2020**, *7*, 205.

(2) Anuradha, J. et al. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science* **2015**, *48*, 319–324.

(3) Uddin, M. F.; Gupta, N., et al. In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education*, 2014, pp 1–5.

Declaration

- 1. I understand what plagiarism is and am aware of the University's policy in this regard.
- 2. I declare that this REPORT (e.g. essay, report, project, assignment, dissertation, thesis, etc.) is my own original work. Where other people's work has been used (either from a printed source, internet or any other source), this has been properly acknowledged and referenced in accordance with the requirements as stated in the University's plagiarism prevention policy.

Signature
Yasekwa Dutywa 18 August 2023