KNOWLEDGE GRAPH

Raima Roj 222BDA10

Arya S 222BDA40

Ann Maria Joy 222BDA47

Hari Krishna 222BDA57

Department of Advanced Computing, St. Joseph's University

BD2P1: Foundations of Data Science Lab

Dr. Jayathi Bhadra

AIM

Our aim is to build a Knowledge Graph using sample data. A knowledge graph is a type of graph database that is designed to store and represent knowledge in a structured form. Unlike traditional databases that organize data in tables with rows and columns, knowledge graphs use a graph structure with nodes and edges to represent information as entities and their relationships which helps us in: Better Data integration, more flexibility in querying, Improved data governance, the key highlight is improved Data discovery and more sophisticated reasoning and analysis. In this project we are first scraping the data and saving it in a csv format, before we get into performing operations on the data we install the required packages such as Spacy, Wikipedia-API to scrape the data from Wikipedia and further we use the dataset obtained and run operations to extract the entities first and then its respective relationship between the entities and finally we represent them in the form of nodes(which contains the Entities) and vertices(which represents the relation) Thus building the Knowledge graph.

LITERATURE REVIEW

Knowledge graphs have become a popular area of interest in recent years, due to their ability to represent complex relationships between data points in a structured form. When we looked upon many research papers and projects, we came across similar ways to implement this project, but it was on different datasets, the datasets they performed on included MS Dhoni, Virat Kohli, Covid-19 etc. In our project, we used a simple dataset that

included details about songs, lyricists, artists, etc. to create a knowledge graph based on them.

Web scraping is the process of extracting data from websites automatically. It has been widely used in various fields such as data mining, business intelligence, and research. Web scraping can be performed using programming languages such as Python, Java, and Ruby. There are various tools and libraries available for web scraping, including Beautiful Soup, Scrapy, and Selenium. Web scraping has some ethical and legal concerns, as it can violate website terms of service and copyright laws. Therefore, it is important to ensure that the data being scraped is legal and ethical.

Knowledge graphs:

Knowledge graphs are a way of representing knowledge and relationships between entities in a graph format. They have been used in various fields such as natural language processing, information retrieval, and semantic search. Knowledge graphs can be created using various knowledge representation formats, such as RDF, OWL, and Neo4j.

Knowledge graphs have some advantages over other knowledge representation methods, such as hierarchical taxonomies and relational databases. They can capture complex relationships between entities and support flexible querying and inference.

Machine learning and knowledge graphs:

Machine learning techniques can be used to analyze knowledge graphs and make predictions or extract insights. Graph-based algorithms, such as PageRank and centrality measures, can be used to identify important nodes in the graph. Natural language processing techniques, such as

named entity recognition and relation extraction, can be used to extract information from unstructured text and add it to the knowledge graph. Various studies have shown the effectiveness of using knowledge graphs for machine learning tasks. For example, a study by Goyal et al. (2018) used a knowledge graph to improve the accuracy of entity linking in a natural language processing system.

Web scraping and knowledge graphs are powerful tools for extracting and representing knowledge from the web. They can be used in various machine learning tasks, such as natural language processing and information retrieval. However, it is important to consider ethical and legal concerns when using web scraping, and to carefully design the knowledge graph to capture the relevant relationships between entities.

METHODOLOGY

To build knowledge graphs from text it is important to help our machine understand natural language and to perform this we use certain NLP techniques such as sentence segmentation, entity extraction and relation extraction.

A knowledge graph is a structured data representation that encapsulates links between knowledge in a certain topic. Information is represented by nodes (informational entities) and edges (relationships between those entities) in this sort of graph database. A knowledge graph can be used to model complicated systems and conduct data analysis since the nodes and edges in it represent real-world concepts and their interactions. For data integration, information retrieval, and reasoning, knowledge graphs are frequently employed in artificial intelligence and machine learning applications. They are also utilized in a variety of industries, including e-

commerce, healthcare, and finance, for a variety of purposes including knowledge management, fraud detection, and recommendation systems. Knowledge graphs are often composed of datasets from several sources, many of which have different structural characteristics. Together, schemas, identities, and context give several types of data structure. Identity categories the underlying nodes appropriately, schemas give the foundation for the knowledge graph, and context establishes the environment in which that knowledge is found. Words having many meanings can be distinguished using these elements. This makes it possible for tools to distinguish between Apple, the brand, and apple, the fruit, like Google's search engine algorithm.

Natural language processing (NLP):

It is used by machine learning-powered knowledge graphs to generate a comprehensive representation of nodes, edges, and labels through a procedure termed semantic enrichment. Knowledge graphs can recognise specific objects and comprehend the links between various objects when data is fed through this procedure. Then, more datasets that are pertinent and have a comparable character are compared to this working knowledge and merged into them. When a knowledge graph is finished, it enables search and question-answering systems to retrieve and reuse thorough responses to specific queries. The same tools can be used in a business setting, removing the need for human data collecting and integration effort to assist business decision-making, just as they can be used to consumer-facing products that show their capacity to save time.

By drawing links between data points that might not have been obvious before, the data integration efforts centered on knowledge graphs can also aid in the production of new knowledge.

Sentence Segmentation:

It is the process of dividing a text document into individual sentences. In the context of a knowledge graph, sentence segmentation can be used to extract information and relationships between entities from natural language text.

One approach to sentence segmentation in a knowledge graph is to use natural language processing (NLP) techniques, such as dependency parsing and part-of-speech tagging, to identify the boundaries between sentences. These techniques can analyze the grammatical structure of the text to determine where one sentence ends and the next one begins.

Once the text has been segmented into sentences, the information and relationships between entities can be extracted using entity extraction and relationship extraction techniques. These techniques can identify, and extract named entities, such as people, organizations, and locations, and the relationships between them, such as ownership, membership, and location.

Entity Extraction:

Entity extraction, also known as named entity recognition (NER), is a natural language processing technique that involves identifying and extracting important entities or objects from unstructured text data. These entities may include proper nouns, such as people, places, organizations, or specific products, or they may be more abstract concepts, such as dates, times, or numerical values.

The process of entity extraction involves analyzing the text data to identify and classify the entities. This is typically done using machine learning algorithms and techniques, such as statistical models, deep learning, or rule-based methods. These algorithms can be trained on large

datasets of labeled text to learn patterns and features that are indicative of specific entity types.

Entity extraction is widely used in a variety of applications, including information retrieval, search engines, chatbots, sentiment analysis, and text classification. By identifying important entities in the text, it can help to improve the accuracy and relevance of these applications. Additionally, entity extraction can help to automate the process of data entry and analysis, making it easier and more efficient to work with large volumes of text data.

Relation Extraction:

Relation extraction plays a crucial role in building a knowledge graph. A knowledge graph is a type of knowledge representation that stores information in the form of nodes and edges. Nodes represent entities, while edges represent the relationships between those entities.

To build a knowledge graph, relation extraction is used to identify and extract the relationships between entities mentioned in a text. This involves analyzing the text to identify the entities mentioned and the type of relationship that exists between them. For example, in the sentence "Barack Obama was the 44th President of the United States," the entities mentioned are "Barack Obama" and "United States," and the relationship between them is "President of."

Once the relationships have been extracted, they can be used to build a graph that represents the knowledge contained in the text. Each entity is represented as a node in the graph, and each relationship is represented as an edge connecting the nodes.

IMPLEMENTATION

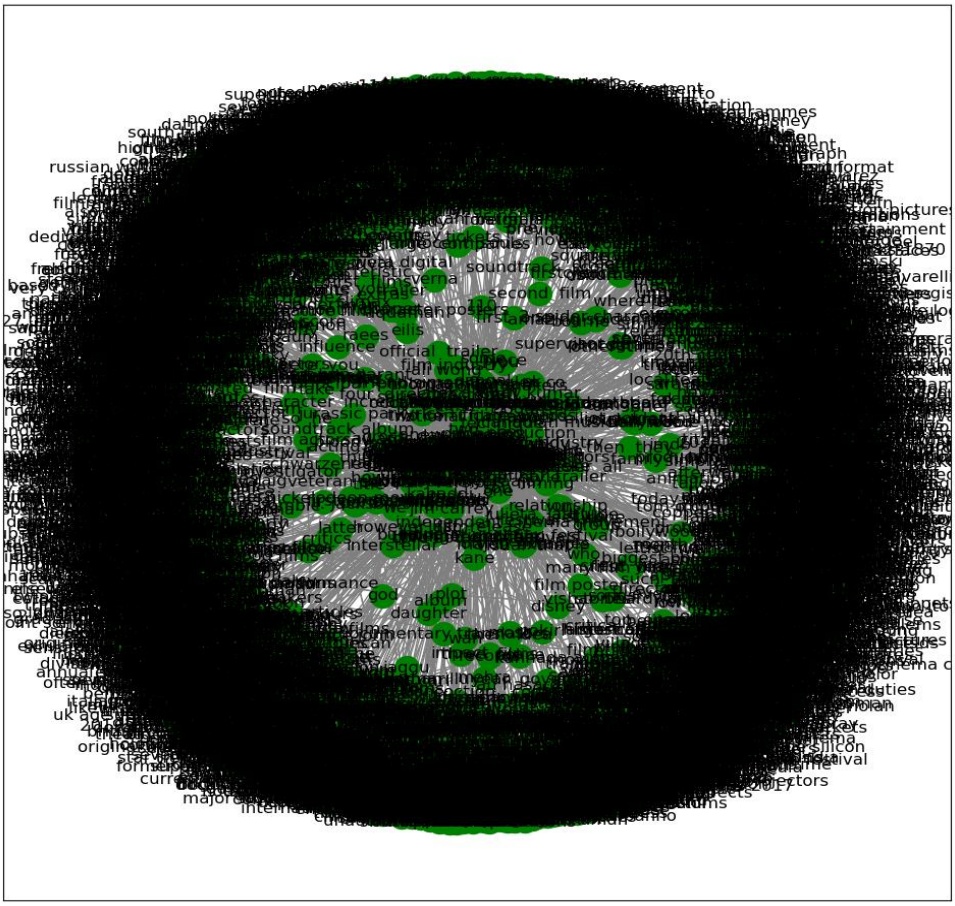
In our project we performed a few steps which includes:

1. Installed the dependencies and scraped the data and saved it in a csv format.
2. Loaded the dataset.
3. To build the knowledge graph, we had to split the text document into sentences. This step is called sentence segmentation which is done using NLP.
4. The nodes and the edges connecting them are crucial components in the construction of a knowledge graph. The entities that appear in Wikipedia sentences will make up these nodes. These entities' connections to one another are represented by edges. Using the sentence structure, we will utilize an unsupervised method to extract these parts. The fundamental concept is to go through a sentence and identify the subject and object as you come across them.
5. There are few groups specified:
    - Group 1: Defined a few empty variables in this.
    - Group 2: Loop through the tokens in the sentence.
    - Group 3: If the token is the subject, then it will be captured as the first entity in the ent1 variable.
    - Group 4: If the token is the object, then it will be captured as the second entity in the ent2 variable.
    - Group 5: Once the subject and the object in the sentence are captured, we will update the previous token and its dependency tag.

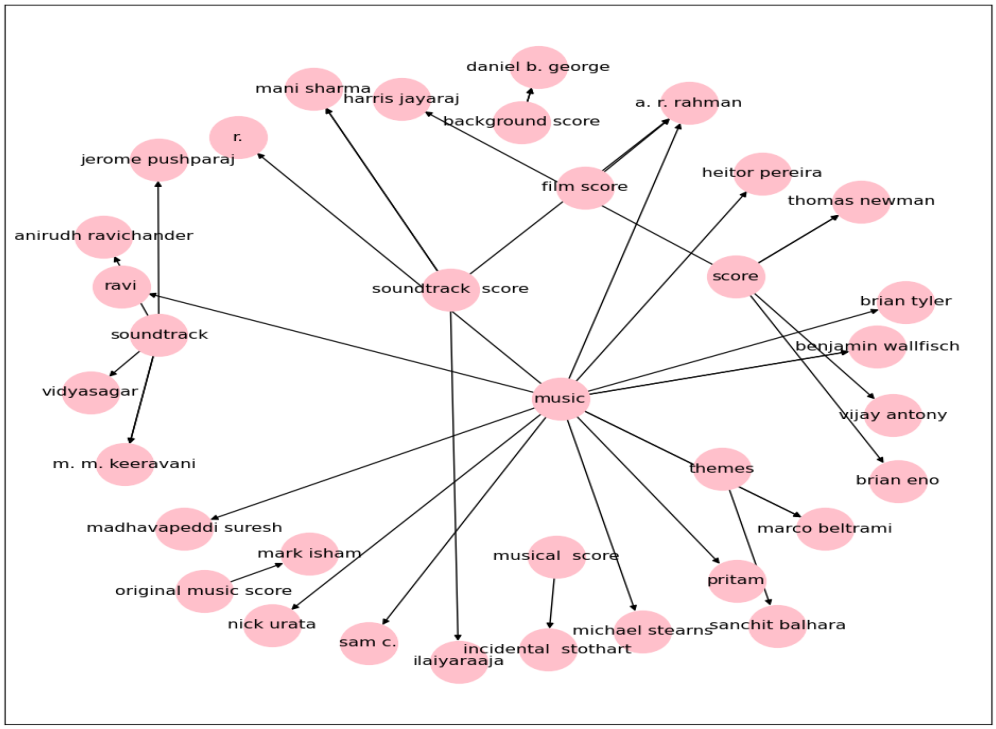
6. We require edges to link the nodes (entities) together to construct a knowledge graph. The connections between two nodes are represented by these edges. The idea that we have is that the primary verb in a phrase serves as the predicate.
7. We will finally create a knowledge graph from the extracted entities (subject-object pairs) and the predicates (relation between entities).

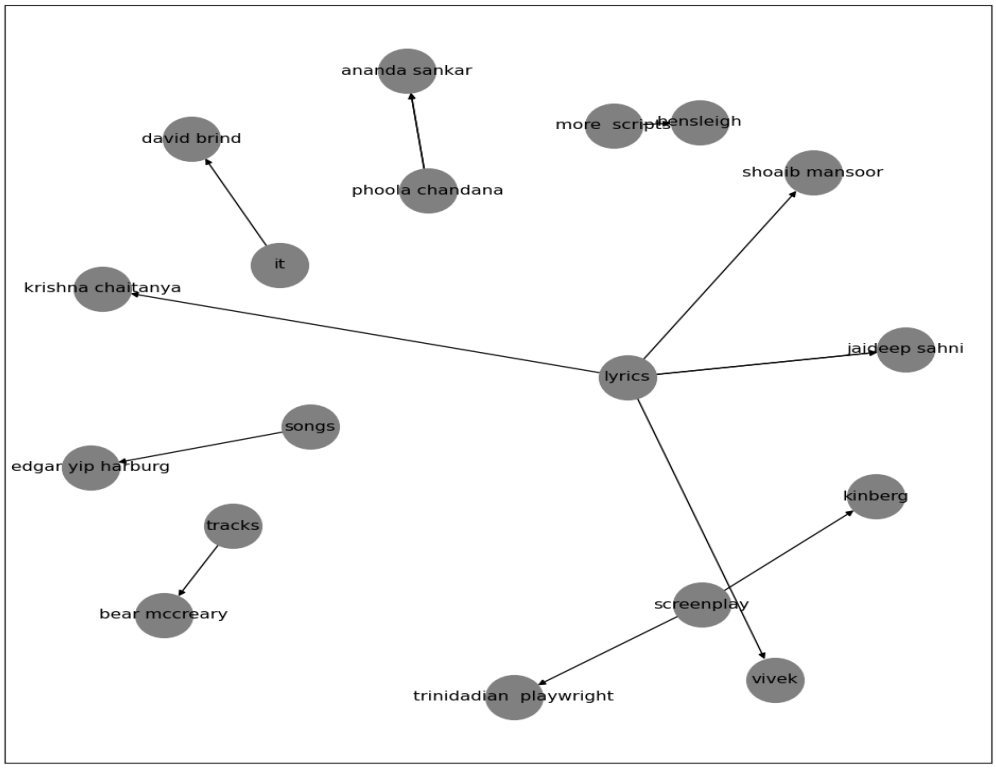
EXPERIMENTS AND RESULTS

The graph below gives the complete details of the text we gave as input.



The graph seemed too complicated to understand therefore we had to select a few entities to understand it better.





CONCLUSION

Knowledge graphs have recently emerged as a powerful tool for representing and organizing information in a structured format, which uses nodes and edges to represent entities and their relationships with other entities. In our project we have used a sample text which includes details of a variety of artists, lyricists, etc. We used their information to create a relation between each of the entities and represent them pictorially.

The use cases of knowledge graph include:

- Analytics Modernization
- Data fabric
- Data lake acceleration
- Operational risk
- Semantic search

REFERENCES

https://www.analyticsvidhya.com/blog/2019/10/how-to-build-knowledge-graph-text-using-spacy/

https://neptune.ai/blog/web-scraping-and-knowledge-graphs-machine-learning

https://hami-asmai.medium.com/relationship-extraction-from-any-web-articles-using-spacy-and-jupyter-notebook-in-6-steps-4444ee68763f