**Semantic Dataset Analysis**

**Phase # 02**

**Group Members: Roll Numbers:**

|  |  |  |
| --- | --- | --- |
| 1 | Shazia Nazir | NUM-BSCS-2022-09 |
| 2 | Ayesha Tahir | NUM-BSCS-2022-17 |
|  |  |  |

**Instructor:** Sir Muhammad Bilal

**Course:** Data Analysis and Visualization



**Department of Computer Sciences**

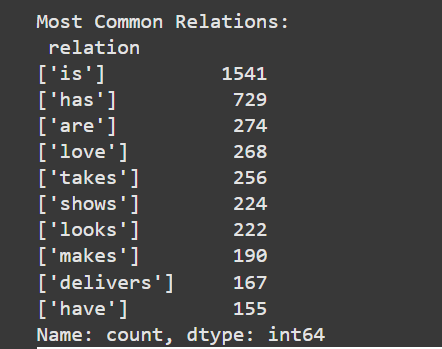
**Namal University Mianwali, Pakistan**

## Task 1: Exploratory Data Analysis (EDA)

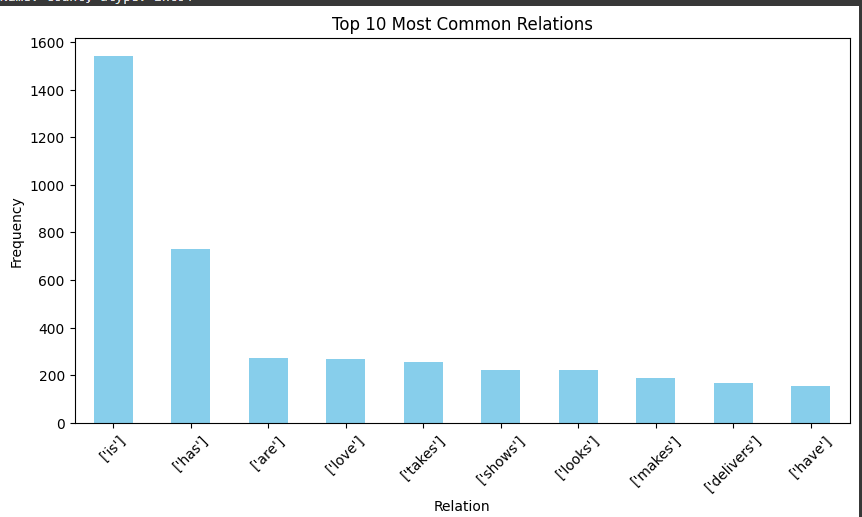
The EDA performs an in-depth exploratory data analysis on a knowledge graph dataset (kG nlp.csv). The goal is to understand entity relationships through visual and statistical techniques.

### Visualization:

***Non-Graphical visualization***

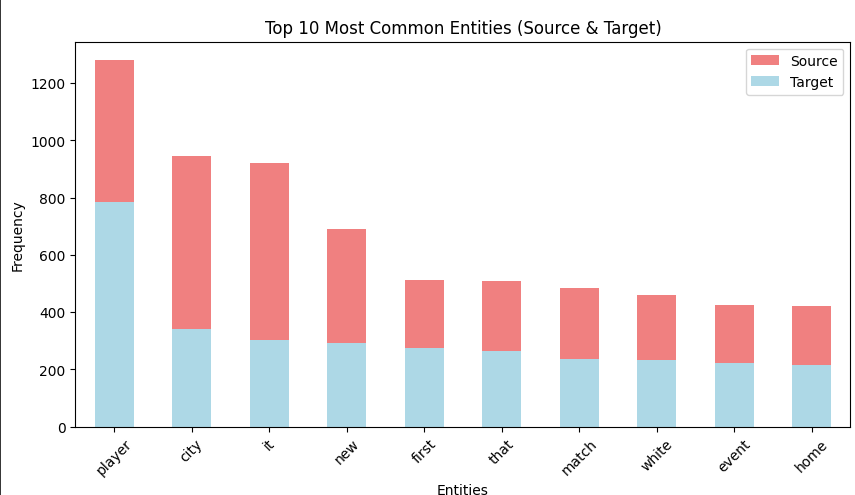


***Graphical visualization***



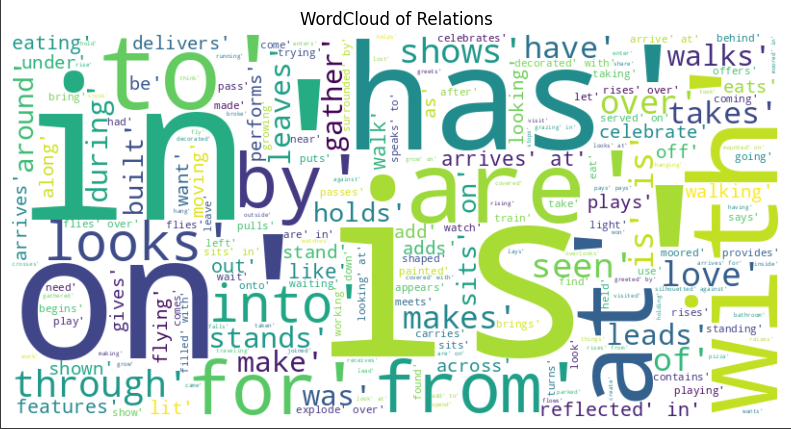
***Source and Target Frequency Relationship***

Frequency counts of the most common 'relation', 'source', and 'target' columns are calculated.



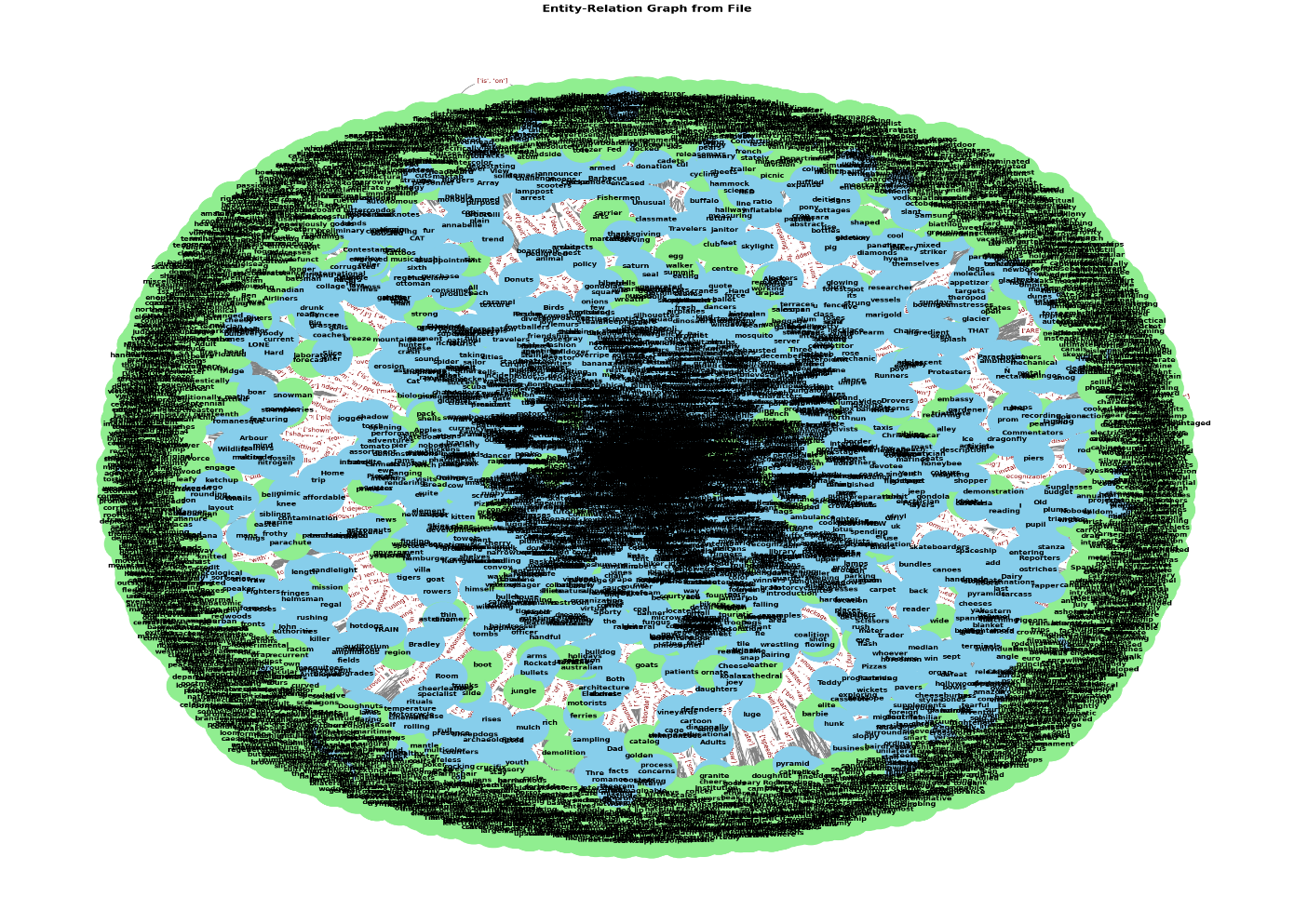
Two bar plots are created to visualize the top 10 most frequent relations and entities in the dataset using matplotlib.

***WordCloud of Relationship***



## **Task 2: Entity-Relation Graph Visualization**

This task visualizes the structure of relationships between entities using a **Directed Graph (DiGraph)** from the dataset kG nlp.csv. Each entity in the dataset acts as a **node**, and the **relations** act as **edges** connecting those nodes.

****

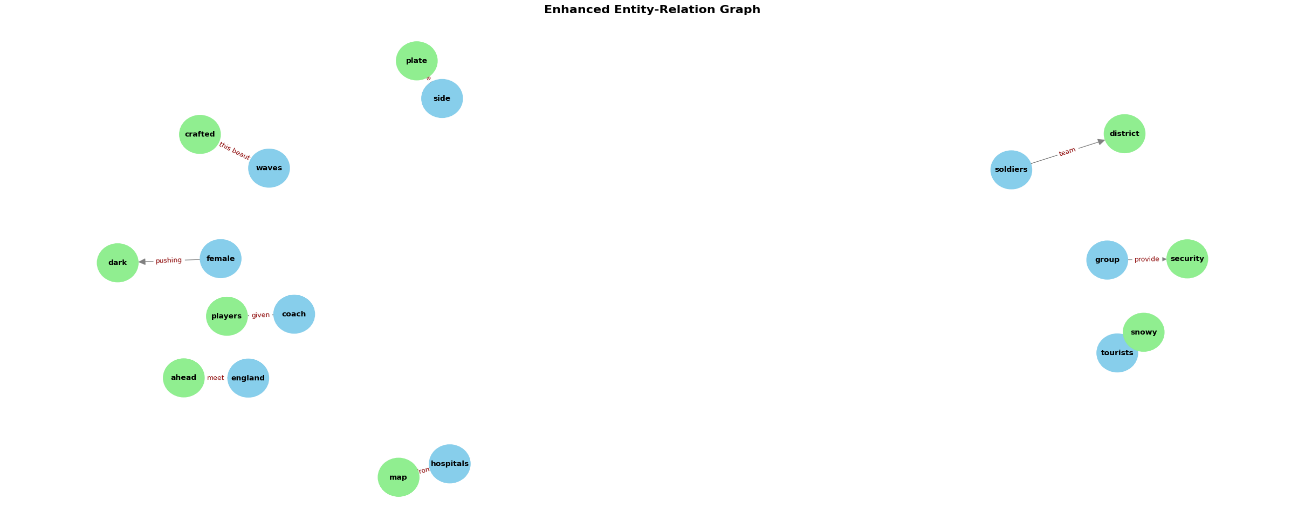
**Nodes** are categorized as source and target, colored in **sky blue** and **light green** respectively.

**Interpretation of the Graph:**

* The central dense cluster represents entities with a **high degree of connectivity**—commonly appearing as either sources or targets in multiple relations.
* Entities on the outer edge are **less connected**, often playing a role in fewer relationships.
* The visual provides **insight into semantic density**, helping identify dominant entities and the richness of their interconnections.
* It’s a helpful tool for **exploring knowledge graphs** and for understanding how entities influence or relate to one another in textual datasets.

## **Task 3: Enhanced Entity-Relation Graph Visualization**

This version focuses on a smaller sample of entity relationships for better readability and interpretability.

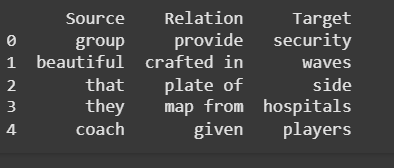


***Example insights from the graph:***

* "group" → "security" labeled "provide" shows a semantic connection suggesting functional roles.
* "coach" → "players" labeled "given" represents a contextual relationship in sports

**Task 4: Triplet Extraction**

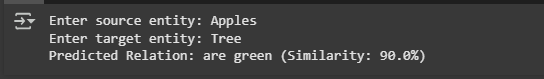
Triplet Extraction is used to convert raw dataset entries into this clear format so that relationships between entities can be easily visualized, analyzed, or predicted using tools like graphs and machine learning.



This task reads the original dataset and extracts meaningful triplets in the form of (Source, Relation, Target).

## Task 5: Predict Most Relevant Relation

This Python script loads a dataset of extracted triplets (source, relation, target) and uses fuzzy string matching (via the fuzzywuzzy library) to predict the most likely relation between any two user-provided entities.



## Conclusion:

In this phase, we explored a knowledge graph dataset to understand how different entities are connected through various relationships. Using both visual and non-visual methods, we analyzed the frequency and structure of these connections. Graph visualizations helped us identify important entities and how they interact with others. Finally, we implemented a simple prediction model using fuzzy matching to guess the most relevant relation between any two entities. Overall, this analysis gave us meaningful insights into the structure and semantics of the dataset.