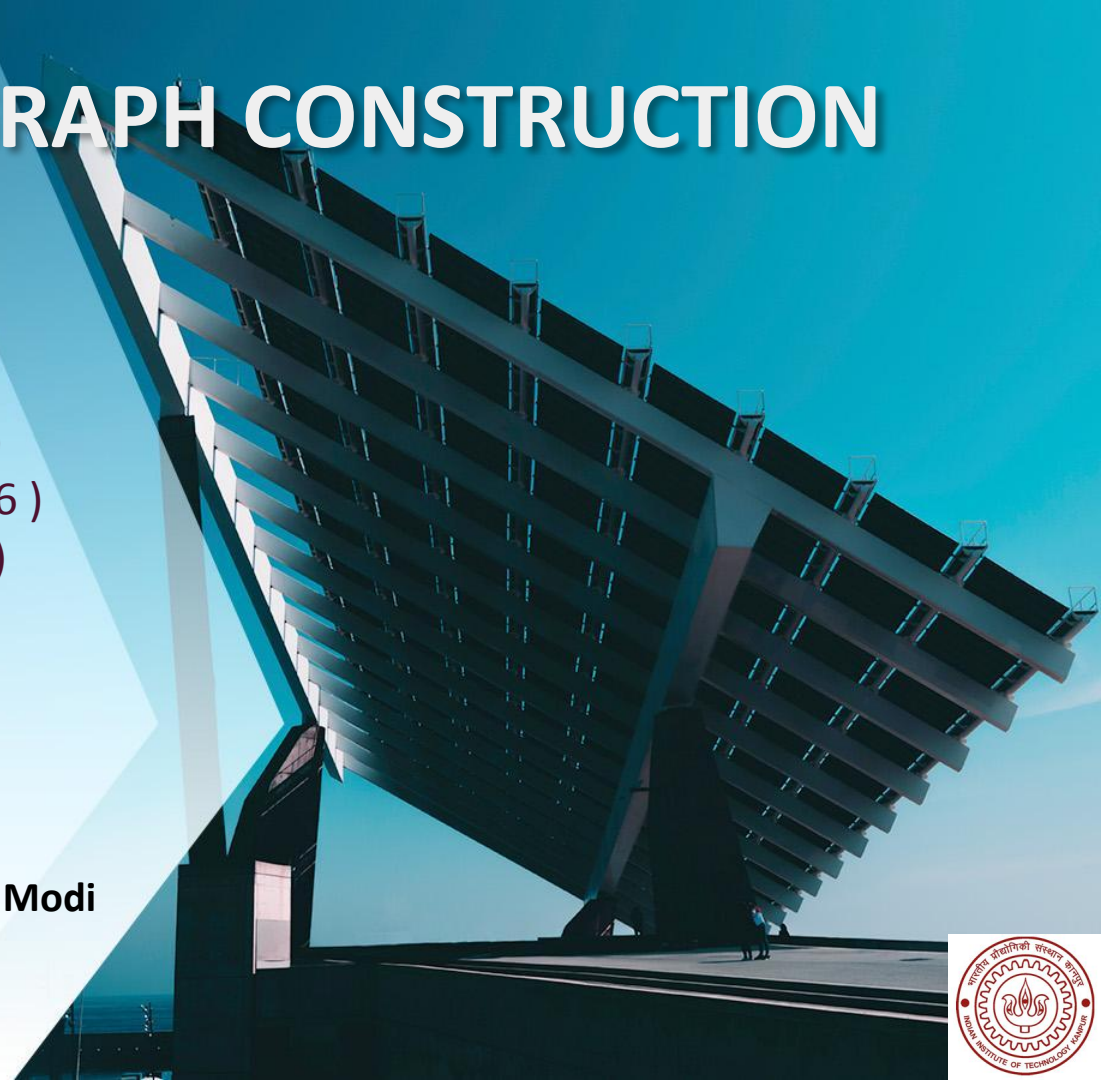
KNOWLEDGE GRAPH CONSTRUCTION



- **Project Number** : 18
- **Team Members** :

Yash Kumar (19211274)
Nitesh Trivedi (19211266)
Vishal Singh (19211273)

Supervised by : - **Asst. Prof. Ashutosh Modi**

KNOWLEDGE GRAPH

- The **Knowledge Graph** is a **knowledge** base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources

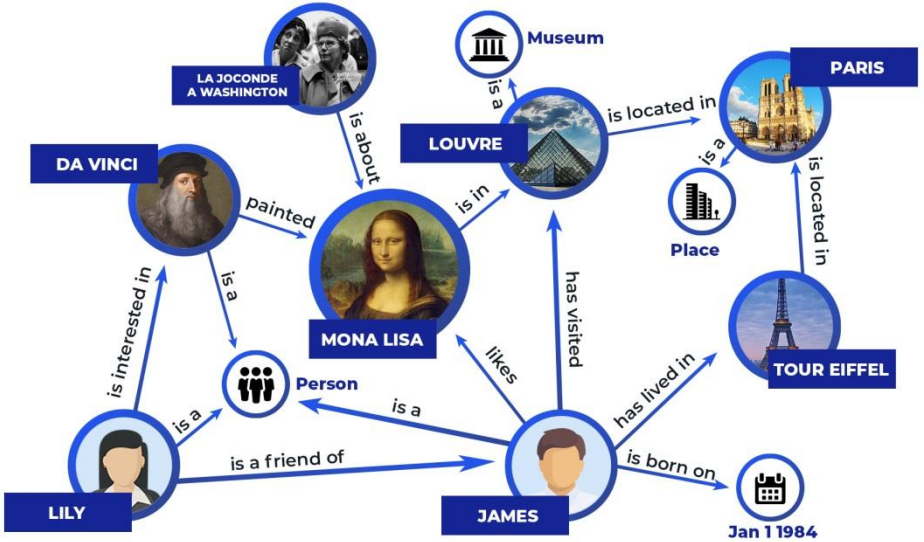


image by https://yashuseth.files.wordpress.com/2019/10/knowledge-graph.jpg
definition by https://en.wikipedia.org/wiki/Knowledge_Graph

PROJECT DESCRIPTION

Motivation : **Knowledge Graph** (KG) plays a crucial role in many modern applications like searching, question answering, data integration, recommendation systems etc., across several domains such as healthcare, geosciences, manufacturing, aviation, power, oil and gas. We plan to construct an isolated **Knowledge Base** for IIT Kanpur website which can further be used for searching.

Problem Statement : Understanding the process of Knowledge Graph Construction, preparing a pipeline and integrating the same for constructing a Knowledge graph for IIT Kanpur website.

LITERATURE REVIEW

Raghunathan, Karthik, et al. "A multi-pass sieve for coreference resolution." ***Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*****. Association for Computational Linguistics, 2010.**

- A simple deterministic approach to coreference resolution.
- Incorporates document level information to be processed at several phases.
- First process mentions available in text and further apply different models in various phases.

Pro

New features or models can be inserted in the system (modular architecture).

Con

Requires very high precision features.

LITERATURE REVIEW

Rusu, Delia, et al. "Triplet extraction from sentences." ***Proceedings of the 10th International Multiconference" Information Society-IS*****. 2007.**

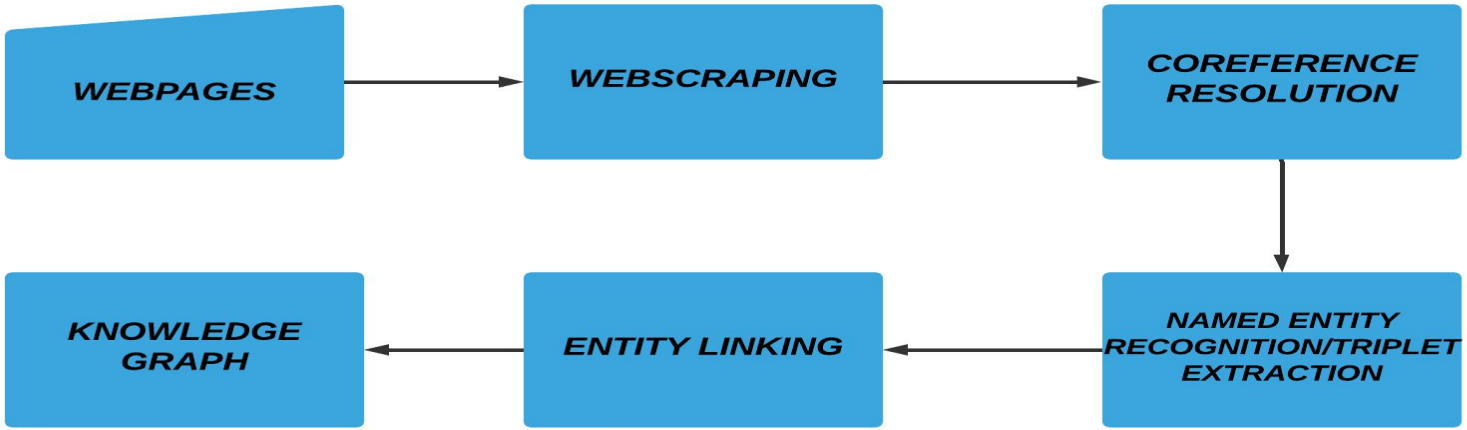
- Well known english syntactic parsers are used to extract subject-predicate-object triplets.
- First parse tree is generated for the sentences, followed by extraction of triplets.
- Four well known parsers are:-
  - TREEBANK PARSERS
  - STANFORD PARSER AND OPEN NLP
  - LINK PARSER
  - MINIPAR

LITERATURE REVIEW

Shen, Wei, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." ***IEEE Transactions on Knowledge and Data Engineering*** **27.2 (2014): 443-460.**

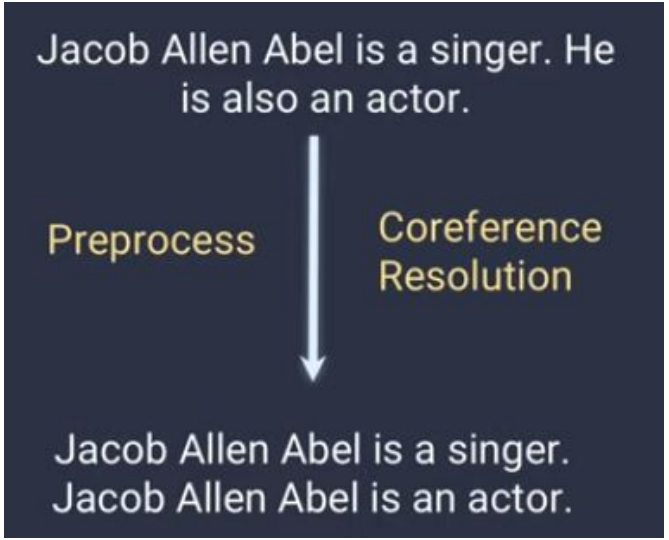
- Authors proposed methods to disambiguate entities during creation of knowledge graph.
- A Entity Linking System Module to link text entities to their corresponding entities in Knowledge Graph .
- Entity linking system consist of the following three modules:
  - Candidate Entity Generation(to filter out irrelevant entities)
  - Candidate Entity Ranking(to find most relevant candidate)
  - Unlinkable Mention Prediction(to check whether selected candidate entity is correct or not)

KG CONSTRUCTION PIPELINE



COREFERENCE RESOLUTION

- **Coreference resolution** is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.
- Technique Used : Stanford CoreNLP
- Challenges:
  - Number and Gender
  - Definiteness of mentions
  - Pronoun Resolution
  - Anaphoricity Determination



definition by https://nlp.stanford.edu/projects/coref.shtml
image by https://www.youtube.com/watch?v=jOugBeAST_Q

NAMED ENTITY RECOGNITION

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entity mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Tasks :

- Extract Names
- Classify the extracted names
- Eg : { '1970s': 'DATE', 'Aaron': 'PERSON', 'Eight': 'CARDINAL' }

definition by https://en.wikipedia.org/wiki/Named-entity_recognition
example by Nagiza F. Samatova, NC State Univ lecture
image by https://www.youtube.com/watch?v=jOugBeAST_Q

TRIPLET EXTRACTION

- According to the approach presented in [1], we define a triplet in a sentence as a relation between subject and object, the relation being the predicate. The aim here is to extract sets of the form {subject, predicate, object} out of syntactically parsed sentences, with four parsers, namely Stanford Parser, OpenNLP, Link Parser and Minipar.
- Technique used : Dependency Parsing
- Eg : SVO extracted from our sample data

```
[(Willie Aames, is, director television producer), (Willie Aames, is, screenwriter), (Philip Abbott, was, character actor), (Victor Aaron Ramirez, was, actor)]
```

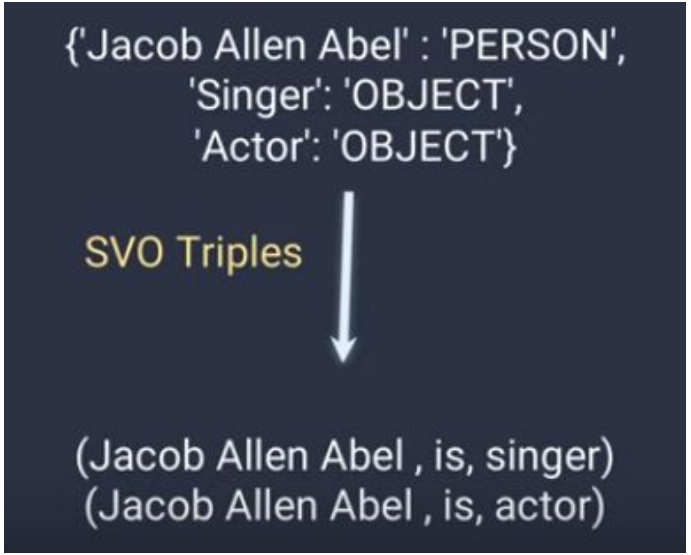


image by https://www.youtube.com/watch?v=jOugBeAST_Q

ENTITY LINKING

- **ENTITY LINKING :**
  - Task to link entity mentions in the text with their corresponding entities in a Knowledge Base.
- **CHALLENGES :**
  - Name variations
  - Entity ambiguity
- **APPROACH :**
  - Candidate Entity Generation.
  - Candidate Entity Ranking.
  - Unlinkable Mention Prediction.
- **CANDIDATE ENTITY GENERATION** :
  - For each entity in text, filter out irrelevant entities from Knowledge Base and retrieve Candidate entity set which contains possible entities.
    - **NAME DICTIONARY BASED APPROACH** : Create a name dictionary D between various names and their possible mapping entities and exploit this constructed name dictionary to generate Candidate entities.

ENTITY LINKING

- **CANDIDATE ENTITY RANKING :**
  - For any entity if |candidate entity| > 1, find that candidate entity which is the most likely for that entity.
  - **Approaches :**
    - Supervised Ranking method:
      - Binary Classification
      - Probabilistic Method
      - Graph based Approach
    - Unsupervised Ranking Method :
      - Vector Space Model(VSM)
      - Information retrieval based method.

- **UNLINKABLE MENTION PREDICTION:**
  - This module to validate whether top ranked entity is the target entity for text entity or not.
  - Using threshold value to validate.
  - If Ranking(Score/Value) of candidate entity is less than threshold then the candidate entity will not be considered as correct candidate.

DESIGNED KNOWLEDGE GRAPH

Actors Dataset scraped from Wikipedia

- Sample data scraped of 25 actors considered for processing.
- ActorsDataSet preprocessed and saved in ActorsPreprocessedDataset(coreference resolved).
- Preprocessed DataSet used to extract SVO triples and NER.
- Then process of Entity Linking to construct the Knowledge Graph Construction.

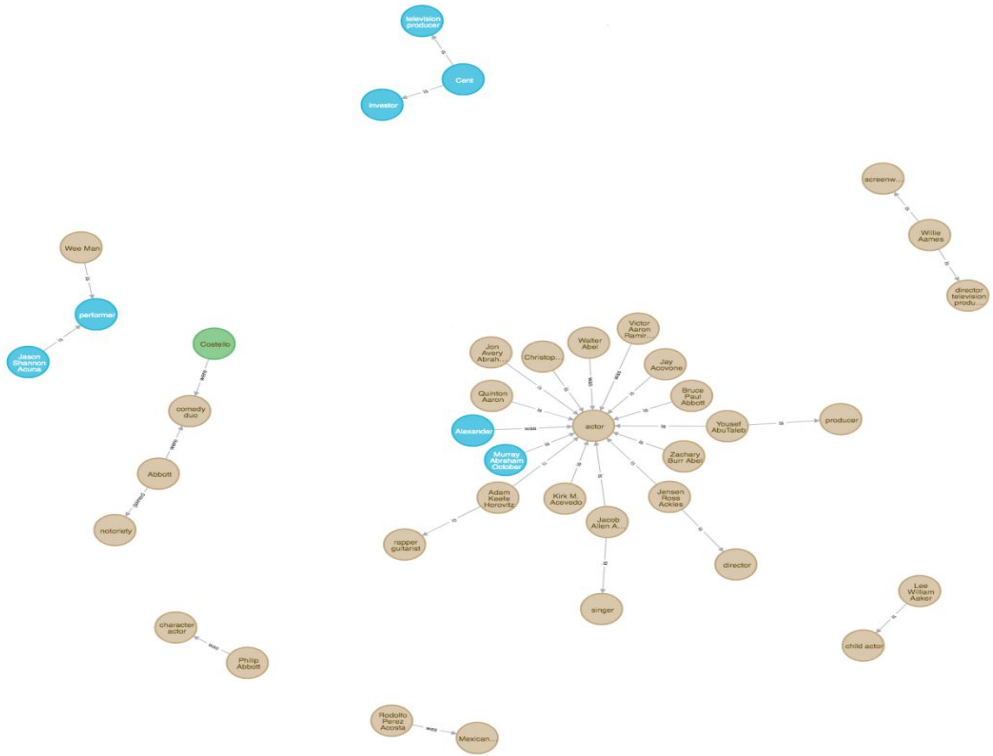


image generated from https://10-0-1-214-34491.neo4jsandbox.com/browser/

FUTURE SCOPE/PHASE 2

- Scraping IIT Kanpur data(ongoing process) :
  - We plan to scrape the website data to run the same via our KG construction pipeline
- Querying the Knowledge Graph constructed for information retrieval
  - The created KG will be queried to get the results which will be compared with Google API
- Comparison with Google KG API results
  - Here, we will compare the results obtained from our KG with Google API results.

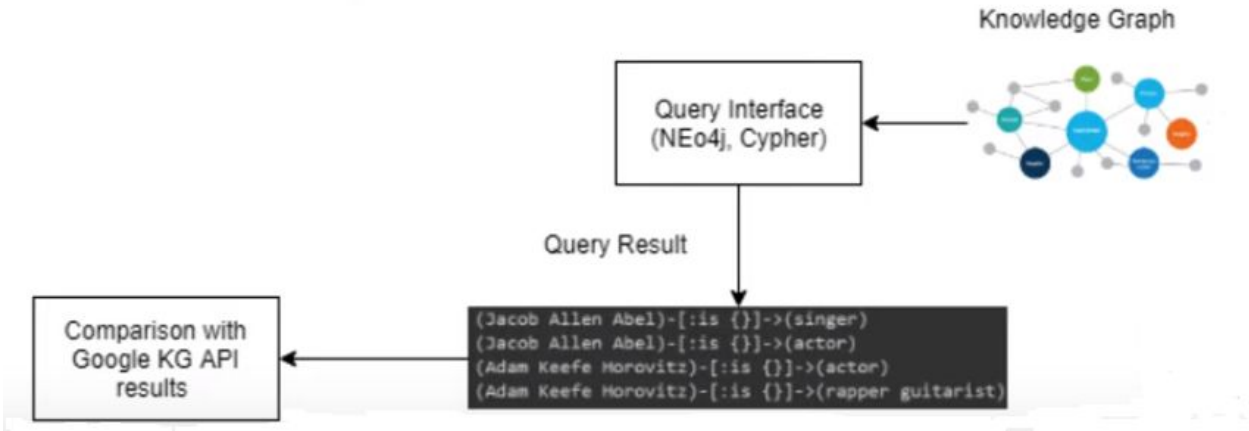


image by https://www.youtube.com/watch?v=jOugBeAST_Q

REFERENCES

[1]-Raghunathan, Karthik, et al. "A multi-pass sieve for coreference resolution." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

[2]-Rusu, Delia, et al. "Triplet extraction from sentences." *Proceedings of the 10th International Multiconference" Information Society-IS*. 2007.

[3]-Shen, Wei, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2014): 443-460.

[4]-Elango, Pradheep. "Coreference resolution: A survey." *University of Wisconsin, Madison, WI* (2005).

[5]-Dali, Lorand, and Blaz Fortuna. "Triplet extraction from sentences using svm." *Proceedings of SiKDD* 2008 (2008).

[6]-https://www.fosteropenscience.eu/content/coreference-resolution-challenges-and-solutions

[7]-https://github.com/KiranMayeeMaddi/NLP