Milestone 1: Mini Knowledge Graph Builder

Infosys Springboard Internship – AI Domain

Submitted by: Meenakshi R
Integrated MSc Computer Science (AI & ML)
Nehru Arts and Science College, Kanhangad

Project Title

Cross Domain Knowledge Mapping

Dataset

The dataset titled "cross domain article.csv" consists of 6,877 records collected from multiple knowledge domains. It contains three columns: category, title, and body. Each entry represents an article or document, where the category specifies the domain (such as Arts Culture, Technology, Health, or Environment), the title gives a brief summary, and the body contains detailed textual content. The dataset is text-rich and unstructured, providing a diverse collection of real-world topics that enable meaningful analysis of entities and relationships across domains. This makes it highly suitable for Natural Language Processing (NLP) tasks such as named entity recognition, relation extraction, and knowledge graph construction.

From the above dataset, a processed version named "tripless.csv" was created for relation extraction and knowledge graph generation. This refined dataset contains 400 records and consists of four columns: Text, Subject, Relation, and Object. Each row represents a meaningful triple derived from a sentence in the original corpus — identifying key entities (Subject and Object) and the Relation connecting them. For example, in a sentence like "Marie Curie discovered Radium," "Marie Curie" serves as the Subject, "discovered" as the Relation, and "Radium" as the Object. This structured representation enables the transformation of unstructured text into a knowledge graph, where entities are visualized as nodes and their semantic connections as edges, providing a clear understanding of relationships across diverse topics.

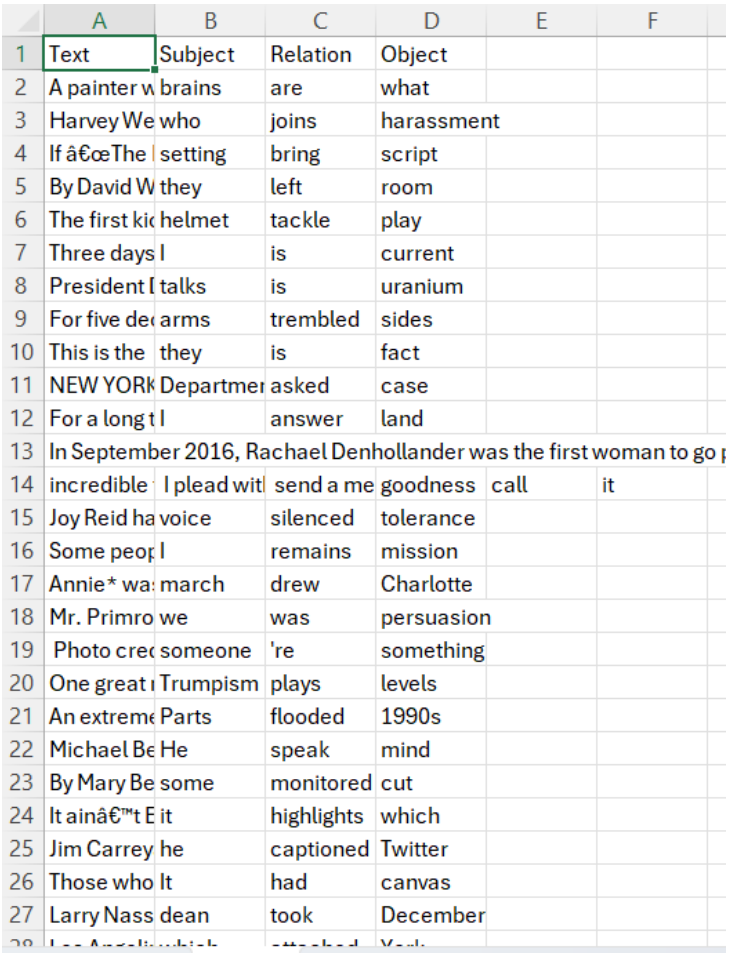
Summary of Extracted Entities and Relationships

From the dataset, a wide variety of entities were extracted, including people, organizations, locations, dates, and artistic or scientific works. Examples include entities such

as Carolyn Kramer (PERSON), The New Yorker (ORG), Donald Trump (PERSON), Barcelona (GPE), and Renaissance (ORG). These entities were identified through Named Entity Recognition (NER) techniques applied on the text corpus.

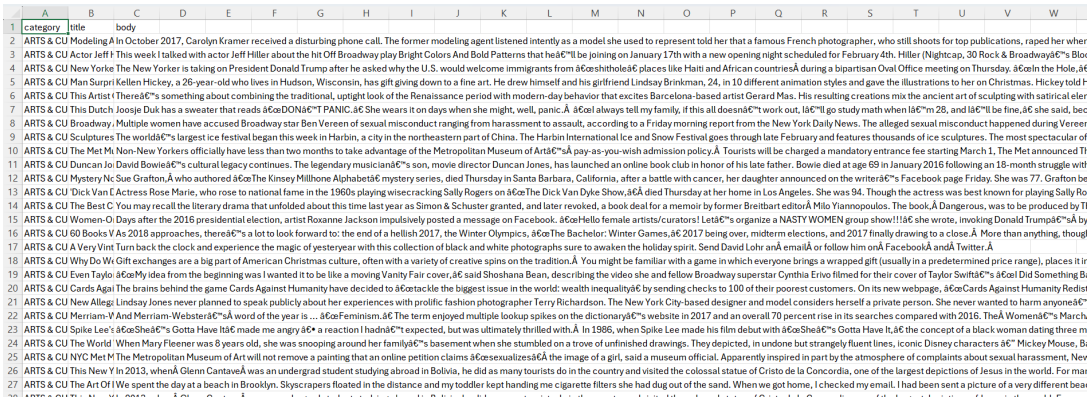
The relationships between these entities were derived using pattern-based extraction methods and semantic role labeling. Common relationships included occupation, office-holder, residence, and work location, representing how one entity connects to another (e.g., Carolyn Kramer → occupation → modeling agent or Gerard Mas → work location → Barcelona).

This structured data of entities and relations was then transformed into a triple format (Subject, Relation, Object) in the tripless.csv file, enabling visualization as a knowledge graph. The resulting graph clearly shows how individuals, organizations, and concepts are interconnected across diverse knowledge domains such as arts, culture, and politics.



	A	B	C	D	E	F
1	Text	Subject	Relation	Object		
2	A painter w	brains	are	what		
3	Harvey We	who	joins	harassment		
4	If â€œThe	setting	bring	script		
5	By David W	they	left	room		
6	The first kic	helmet	tackle	play		
7	Three days	I	is	current		
8	President I	talks	is	uranium		
9	For five dec	arms	trembled	sides		
10	This is the	they	is	fact		
11	NEW YORK	Departmer	asked	case		
12	For a long t	I	answer	land		
13	In September 2016, Rachael Denhollander was the first woman to go					
14	incredible	I plead wit	send a me	goodness	call	it
15	Joy Reid ha	voice	silenced	tolerance		
16	Some peop	I	remains	mission		
17	Annie* wa	march	drew	Charlotte		
18	Mr. Primro	we	was	persuasion		
19	Photo cred	someone	're	something		
20	One great r	Trumpism	plays	levels		
21	An extreme	Parts	flooded	1990s		
22	Michael Be	He	speak	mind		
23	By Mary Be	some	monitored	cut		
24	It ainâ€™t E	it	highlights	which		
25	Jim Carrey	he	captioned	Twitter		
26	Those who	It	had	canvas		
27	Larry Nass	dean	took	December		

tripless.csv



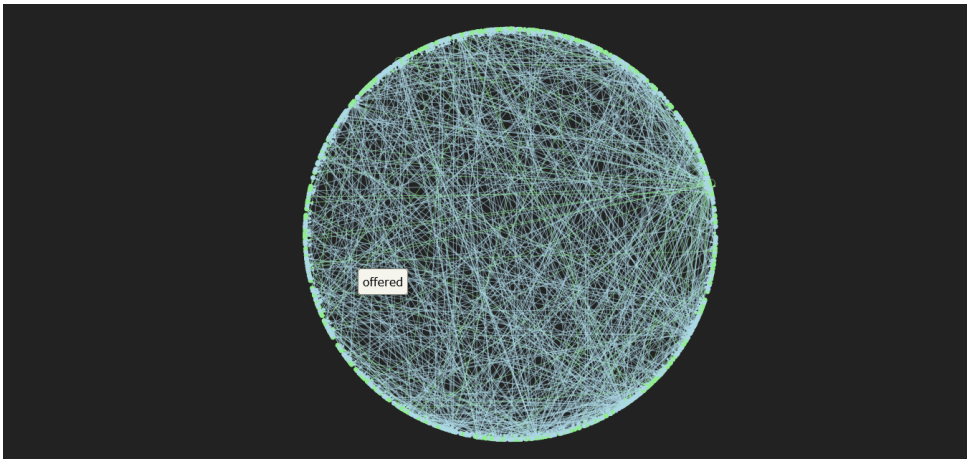
	A	B	C
1	category	title	body
2	ARTS & CU	Modeling Â In October 2017, Carolyn Kramer received a disturbing phone call. The former modeling agent listened intently as a model she used to represent told her that a famous French photographer, who still shoots for top publications, raped her wher
3	ARTS & CU	Actor Jeff H This week I talked with actor Jeff Hiller about the hit Off Broadway play Bright Colors And Bold Patterns that heâ€™ll be joining on January 17th with a new opening night scheduled for February 4th. Hiller (Nightcap, 30 Rock & Broadwayâ€™s Bloc
4	ARTS & CU	New Yorke The New Yorker is taking on President Donald Trump after he asked why the U.S. would welcome immigrants from â€œshitholeâ€ places like Haiti and African countriesÂ during a bipartisan Oval Office meeting on Thursday. â€œIn the Hole,â€
5	ARTS & CU	Man Surpri Kellen Hickey, a 26-year-old who lives in Hudson, Wisconsin, has gift giving down to a fine art. He drew himself and his girlfriend Lindsay Brinkman, 24, in 10 different animation styles and gave the illustrations to her on Christmas. Hickey told H
6	ARTS & CU	This Artist (Thereâ€™s something about combining the traditional, uptight look of the Renaissance period with modern-day behavior that excites Barcelona-based artist Gerard Mas. His resulting creations mix the ancient art of sculpting with satirical eler
7	ARTS & CU	This Dutch Joosje Duk has a sweater that reads â€œDONâ€™T PANIC.â€ She wears it on days when she might, well, panic.Â â€œI always tell my family, if this all doesnâ€™t work out, Iâ€™ll go study math when Iâ€™m 28, and Iâ€™ll be fine,â€ she said, bec
8	ARTS & CU	Broadway Multiple women have accused Broadway star Ben Vereen of sexual misconduct ranging from harassment to assault, according to a Friday morning report from the New York Daily News. The alleged sexual misconduct happened during Vereer
9	ARTS & CU	Sculptures The worldâ€™s largest ice festival began this week in Harbin, a city in the northeastern part of China. The Harbin International Ice and Snow Festival goes through late February and features thousands of ice sculptures. The most spectacular of
10	ARTS & CU	The Met Mu Non-New Yorkers officially have less than two months to take advantage of the Metropolitan Museum of Artâ€™sÂ pay-as-you-wish admission policy.Â Tourists will be charged a mandatory entrance fee starting March 1, The Met announced Th
11	ARTS & CU	Duncan Jo David Bowieâ€™s cultural legacy continues. The legendary musicianâ€™s son, movie director Duncan Jones, has launched an online book club in honor of his late father. Bowie died at age 69 in January 2016 following an 18-month struggle with
12	ARTS & CU	Mystery Nc Sue Grafton,Â who authored â€œThe Kinsey Millhone Alphabetâ€ mystery series, died Thursday in Santa Barbara, California, after a battle with cancer, her daughter announced on the writerâ€™s Facebook page Friday. She was 77. Grafton be
13	ARTS & CU	'Dick Van [Actress Rose Marie, who rose to national fame in the 1960s playing wisecracking Sally Rogers on â€œThe Dick Van Dyke Show,â€Â died Thursday at her home in Los Angeles. She was 94. Though the actress was best known for playing Sally Ro
14	ARTS & CU	The Best C You may recall the literary drama that unfolded about this time last year as Simon & Schuster granted, and later revoked, a book deal for a memoir by former Breitbart editorÂ Milo Yiannopoulos. The book,Â Dangerous, was to be produced by Th
15	ARTS & CU	Women-O Days after the 2016 presidential election, artist Roxanne Jackson impulsively posted a message on Facebook. â€œHello female artists/curators! Letâ€™s organize a NASTY WOMEN group show!!!â€ she wrote, invoking Donald Trumpâ€™sÂ by
16	ARTS & CU	60 Books V As 2018 approaches, thereâ€™s a lot to look forward to: the end of a hellish 2017, the Winter Olympics, â€œThe Bachelor: Winter Games,â€ 2017 being over, midterm elections, and 2017 finally drawing to a close.Â More than anything, thoug
17	ARTS & CU	A Very Vint Turn back the clock and experience the magic of yesteryear with this collection of black and white photographs sure to awaken the holiday spirit. Send David Lohr anÂ emailÂ or follow him onÂ FacebookÂ andÂ Twitter.Â
18	ARTS & CU	Why Do We Gift exchanges are a big part of American Christmas culture, often with a variety of creative spins on the tradition.Â You might be familiar with a game in which everyone brings a wrapped gift (usually in a predetermined price range), places it ir
19	ARTS & CU	Even Taylo â€œMy idea from the beginning was I wanted it to be like a moving Vanity Fair cover,â€ said Shoshana Bean, describing the video she and fellow Broadway superstar Cynthia Erivo filmed for their cover of Taylor Swiftâ€™s â€œI Did Something Ba
20	ARTS & CU	Cards Agai The brains behind the game Cards Against Humanity have decided to â€œtackle the biggest issue in the world: wealth inequalityâ€ by sending checks to 100 of their poorest customers. On its new webpage, â€œCards Against Humanity Redist
21	ARTS & CU	New Allegi Lindsay Jones never planned to speak publicly about her experiences with prolific fashion photographer Terry Richardson. The New York City-based designer and model considers herself a private person. She never wanted to harm anyoneâ€"
22	ARTS & CU	Merriam-W And Merriam-Websterâ€™sÂ word of the year is ... â€œFeminism.â€ The term enjoyed multiple lookup spikes on the dictionaryâ€™s website in 2017 and an overall 70 percent rise in its searches compared with 2016. TheÂ Womenâ€™s March
23	ARTS & CU	Spike Lee's â€œSheâ€™s Gotta Have Itâ€ made me angry â€• a reaction I hadnâ€™t expected, but was ultimately thrilled with.Â In 1986, when Spike Lee made his film debut with â€œShe's Gotta Have It,â€ the concept of a black woman dating three m
24	ARTS & CU	The World When Mary Fleener was 8 years old, she was snooping around her familyâ€™s basement when she stumbled on a trove of unfinished drawings. They depicted, in undone but strangely fluent lines, iconic Disney characters â€" Mickey Mouse, Bi
25	ARTS & CU	NYC Met M The Metropolitan Museum of Art will not remove a painting that an online petition claims â€œsexualizesâ€Â the image of a girl, said a museum official. Apparently inspired in part by the atmosphere of complaints about sexual harassment, Nev
26	ARTS & CU	This New Y In 2013, whenÂ Glenn CantaveÂ was an undergrad student studying abroad in Bolivia, he did as many tourists do in the country and visited the colossal statue of Cristo de la Concordia, one of the largest depictions of Jesus in the world. For mar
27	ARTS & CU	The Art Of I We spent the day at a beach in Brooklyn. Skyscrapers floated in the distance and my toddler kept handing me cigarette filters she had dug out of the sand. When we got home, I checked my email. I had been sent a picture of a very different bea

Original Dataset

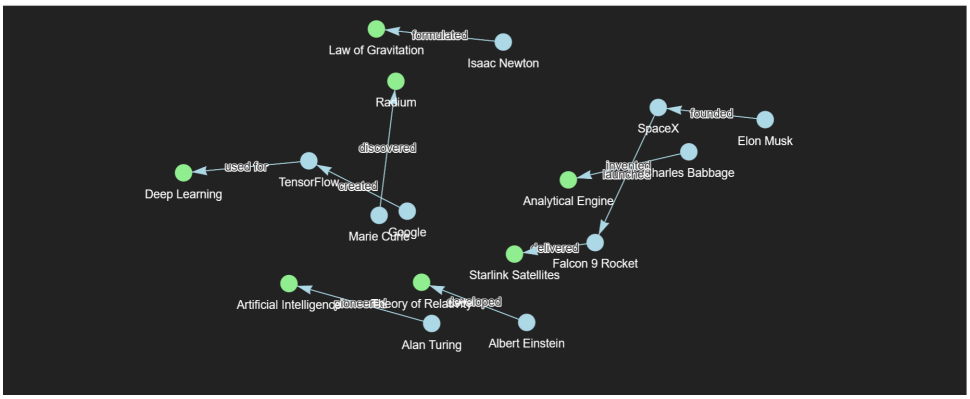
Knowledge Graph Visualization

A knowledge map is a visual representation that organizes and connects information or concepts from different domains, showing how knowledge in one area is related to another. In this project, the knowledge map was developed using Natural Language Processing (NLP) techniques to extract key entities and their relationships from domain-specific textual data. After preprocessing and relation extraction, these entities and relationships were structured into a Knowledge Graph, which served as the foundation of the knowledge map. The map visually links concepts such as technologies, applications, and research areas, enabling the discovery of cross-domain connections. In the internship, I implemented this process by collecting textual datasets, performing NLP-based preprocessing and relation extraction, and constructing an interactive knowledge map using tools like NetworkX and PyVis. This visualization allowed the identification of meaningful relationships between domains, supporting deeper insights and knowledge discovery across multiple fields.

Two knowledge maps were created as part of this milestone. The first knowledge map was generated using the extracted triples from the dataset file tripless.csv, which represents the entities and relationships identified from the collected cross-domain articles. Since this dataset contained a large number of nodes and relationships, the resulting graph appeared congested and densely connected, making it difficult to clearly visualize individual relationships. Therefore, a second knowledge map was also created using a sample dataset provided directly in the code to help better understand and demonstrate the structure, connections, and visualization of entities and relations in a more simplified form. Both maps effectively illustrate how knowledge can be represented through interconnected entities and relationships.



Knowledge Map generated from tripless.csv showing entities and relationships extracted from the cross-domain dataset (graph appears dense due to the large number of nodes and connections).



Simplified Knowledge Map created using sample data directly in the code to clearly visualize entity–relationship connections and understand the structure of the knowledge graph.

Challenges Faced

- Understanding how to structure text data in the form of Subject–Relation–Object triples.
- Getting the visualization to display nodes and edges properly using the PyVis library.
- Setting up and configuring the environment with libraries such as spaCy, pandas, NetworkX, and PyVis took time due to version compatibility issues.
- Visualizing the knowledge graph interactively without overlapping nodes and edges was challenging, as adjusting the graph physics and layout required experimentation.

Reflection – What I Learned from Milestone 1

Through this milestone, I learned how to preprocess unstructured text data using techniques such as tokenization, stop-word removal, stemming, and lemmatization to prepare it for further analysis. I gained hands-on experience in applying Natural Language Processing (NLP) to extract meaningful entities and relationships from text and represent them in the form of subject–relation–object triples. I understood how to organize a structured project environment and manage datasets using pandas. Additionally, I learned how to build and visualize knowledge graphs using NetworkX and PyVis, making it easier to explore the relationships between concepts. This milestone helped me develop a deeper understanding of how AI and NLP can be used to transform raw text into structured, meaningful knowledge representations.