Extracting Graphical Structures from Mixed Data Sources

CMU-Africa Master's Practicum | Spring 2021

Samuel Assefa, Srijan Sood, Parisa Hassanzadeh

AI RESEARCH

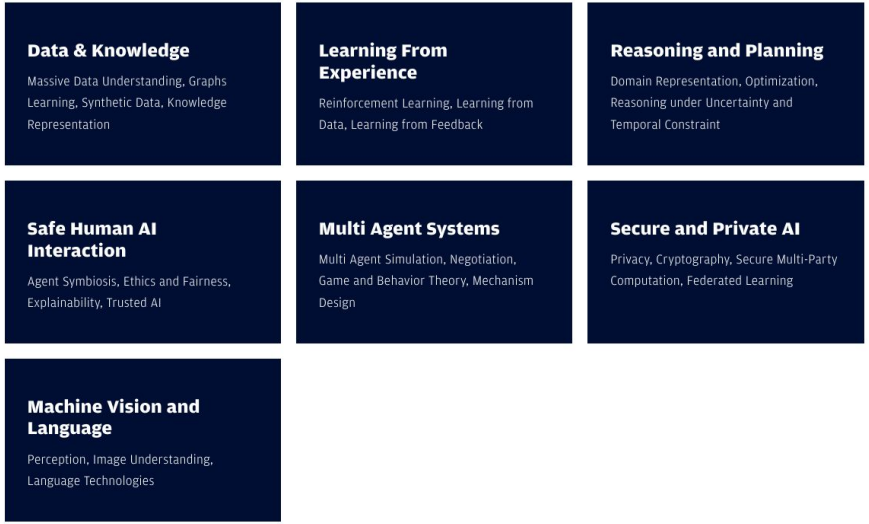
J.P.Morgan

Who are we?

The goal of J.P. Morgan's AI Research program is to explore and advance cutting-edge research in the fields of AI and Machine Learning, as well as related fields like Cryptography, to develop solutions that are most impactful to the firm's clients and businesses.

The AI Research team is headquartered in New York and present in key hubs around the world. Our team is comprised of experts in various fields of AI. They pursue primary research in areas relative to our research pillars as well as concrete problems related to financial services. They partner with various internal teams to accelerate the adoption of AI within the firm. They also work with leading faculty around the world on areas of mutual interest.

Our Research Agenda

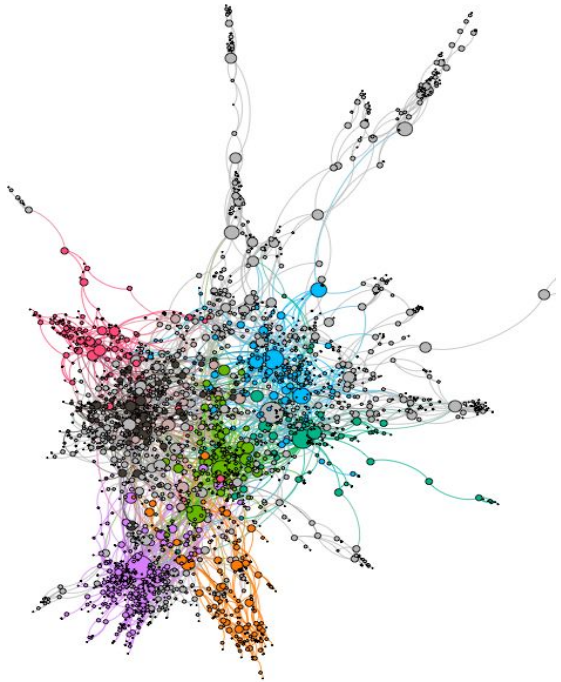


Visit us at https://www.jpmorgan.com/ai.

Project Overview

Data in the Financial Services industry is found in all formats: structured semi-structured and unstructured. However, as data sources vary widely, these datasets are often used independently, and the potential of underlying connections is overlooked. Recent advances in Graph Machine Learning demonstrate the power of leveraging these connections.

This project focuses on forming knowledge graphs by fusing information from multiple data sources and then evaluating the *"goodness"* of the representation. A sample task is forming a graph of publicly traded companies with edges connecting *related* nodes (companies). A knowledge graph of these would capture their relations – their co-occurrences in news articles, their stock price correlations, their business relations (through shared clients/investors/transactions), etc. Community detection on such a graph would clump *similar* companies together (akin to sector-specific indices in the stock market: technology, finance, healthcare, energy, etc.).



Project Data

The team will only be using freely available public data, such as:

- news articles scraped using the New York Times API
- daily stock price data from Yahoo finance
- SEC reports, etc.

The data will be sourced entirely from the web and other open sources using publicly available libraries and APIs.

Outcomes

We expect the team to write a functioning open-source python package AND publish their results. At a minimum, we ask the team to publish their findings as a blog post article/Medium post. As a stretch goal, we would like the team to target publication at a workshop venue of top AI conferences.

We expect the team to produce a weekly summary of progress and deliver a final presentation to our AI Research team. We would like the team to extract graphs from news data, and then evaluate the representational strength of the graphs. One such technique would constitute building a graph using historical news articles (e.g. from 2000 - 2015), and then using it to see if linked entities appear together in the test set corpus (e.g. articles from 2016 - 2020).

Additional Information

Skills required from the team:

- programming skills (python),
- familiarity with data science libraries (pandas, numpy)
- natural language processing knowledge is a plus

Skills acquired by the team:

- data scraping
- natural language processing (e.g. spaCy, hugging-face)
- graph learning (e.g. networx, graphtool)

Supplementary material: An example approach to this project (of extracting a graph from news articles) can be found here.