Extracting Graphical Structures from Mixed Data Sources

MSP PRACTICUM

CLIENT

JP Morgan

TEAM

Aline Niyonsaba

Eric Ngabonzima

Ernest Kufuor

Ryan Harty

ADVISOR

Dr. Moise Busogi

Problem Definition

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 to develop solutions that are most impactful to the firm's clients and businesses.

JPMorgan is looking for a way to identify communities of companies, as well as relationships between companies, without having to guess at them by hand. They would like a more rigorous technical approach to figuring out which companies are related in order to guide processes like investment strategy and fraud detection.

Project Objectives

01

Data Scraping

Scrape financial company data from news, reports and stock market data.

02

Knowledge Graph

Pull entities from unstructured data and store into a knowledge graph.

03

Insights

Identify relationships and gain insights between financial companies through visualizations and graph functions.

04

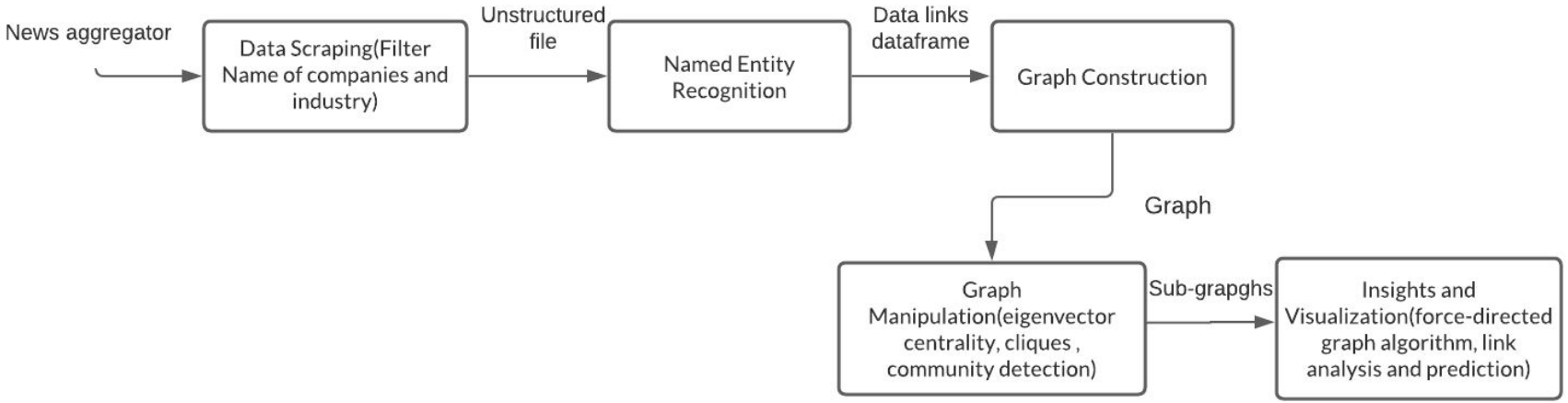
Advance Knowledge

Publicize results to advance knowledge in this field.

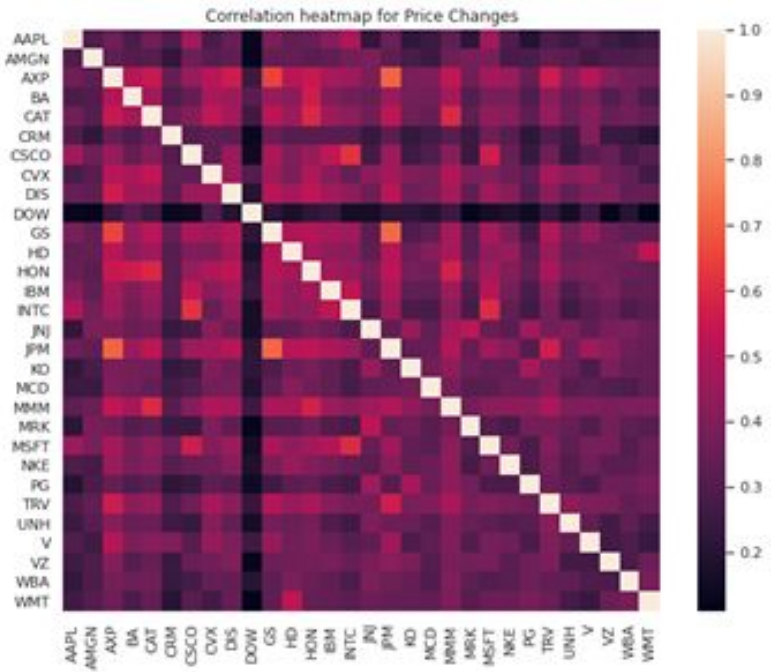
Solution Design

Our system is to be used as a package and used in conjunction with other tools that could aggregate better results.

Block Diagram



Data Sources



	title	media	date	datetime	desc	link	img
0	STUCK IN WALMART PARKING lot after Colorado storm	FOX31 Denver	1 hour ago	2021-03-15 22:26:32.753586	Several drivers were forced into a Walmart par...	https://kdvr.com/news/local/drivers-spend-hour...	
1	Suspect wanted in fatal Monkey Junction Walmar...	Home - WSFX	1 hour ago	2021-03-15 22:26:32.755743	Laron Lee Carter is wanted by the NHC Sheriff'...	https://foxwilmington.com/local-news/suspect-w...	
2	Walmart Theft Investigation Turns Into Chase A...	KVRR	1 hour ago	2021-03-15 22:26:32.757773	— Two people are arrested after a chase where ...	https://www.kvrr.com/2021/03/15/walmart-theft-...	
3	Walmart Calls Ex-Worker's 'Stonewalling' Claim...		1 hour ago	2021-03-15 22:26:32.759685	Law360 (March 15, 2021, 5:49 PM EDT) -- An ex-...	https://www.law360.com/retail/articles/1364916...	
4	Three wanted for questioning in Walmart shopli...	wgxa.tv	1 hour ago	2021-03-15 22:26:32.761621	HOUSTON COUNTY, Ga. -- Perry police are lookin...	https://wgxa.tv/news/local/three-wanted-for-qu...	
...	...	...	...	...	...	...	...
5	Denim Market (COVID-19) to Witness	KSU | The Sentinel Newspaper	10 hours ago	2021-03-15 13:36:50.533737	... to Witness Astonishing Growth by	https://ksusentinel.com/2021/03/15/denim-marke...	

Methods- Named Entity Recognition

```
def get_ner_data(df_row):
    '''
    - function to extract named entities from a paragraph
    - returns two data frames:
        - the first is a dataframe of all unique entities (persons and orgs)
        - the second is the links between the entities
    '''
    paragraph=df_row.content
    #changed above row
    # remove newlines and odd characters
    paragraph = re.sub('\r', '', paragraph)
    paragraph = re.sub('\n', ' ', paragraph)
    paragraph = re.sub("'s", '', paragraph)
    paragraph = re.sub("“", '', paragraph)
    paragraph = re.sub("”", '', paragraph)


    # tokenise sentences
    sentences = tokenize.sent_tokenize(paragraph)
    sentences = [Sentence(sent) for sent in sentences]

    # predict named entities
    for sent in sentences:
        tagger.predict(sent)

    # collect sentence NER's to list of dictionaries
    sent_dicts = [sentence.to_dict(tag_type='ner') for sentence in sentences]

    # collect entities and types
    entities = []
    types = []
    for sent_dict in sent_dicts:
        entities.extend([entity['text'] for entity in sent_dict['entities']])
        types.extend([str(entity['labels'])[1:4] for entity in sent_dict['entities']])
  #The above line is what I changed from the default notebook to get things working

    # create dataframe of entities (nodes)
    df_ner = pd.DataFrame(data={'entity': entities, 'type': types})
    df_ner = df_ner[df_ner['type'].isin(['ORG'])]
    df_ner = df_ner[df_ner['entity'].map(lambda x: isinstance(x, str))]
    df_ner = df_ner[~df_ner['entity'].isin(df_contraptions['contraption'].values)]
    df_ner['entity'] = df_ner['entity'].map(lambda x: x.translate(str.maketrans('', '', string.punctuation)))
```

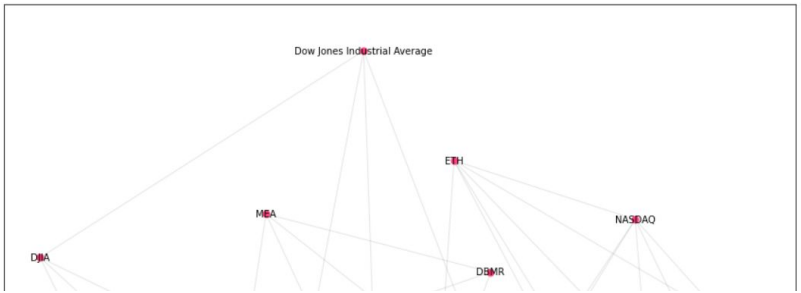
Methods- Graph Creation

4. Plot Edges

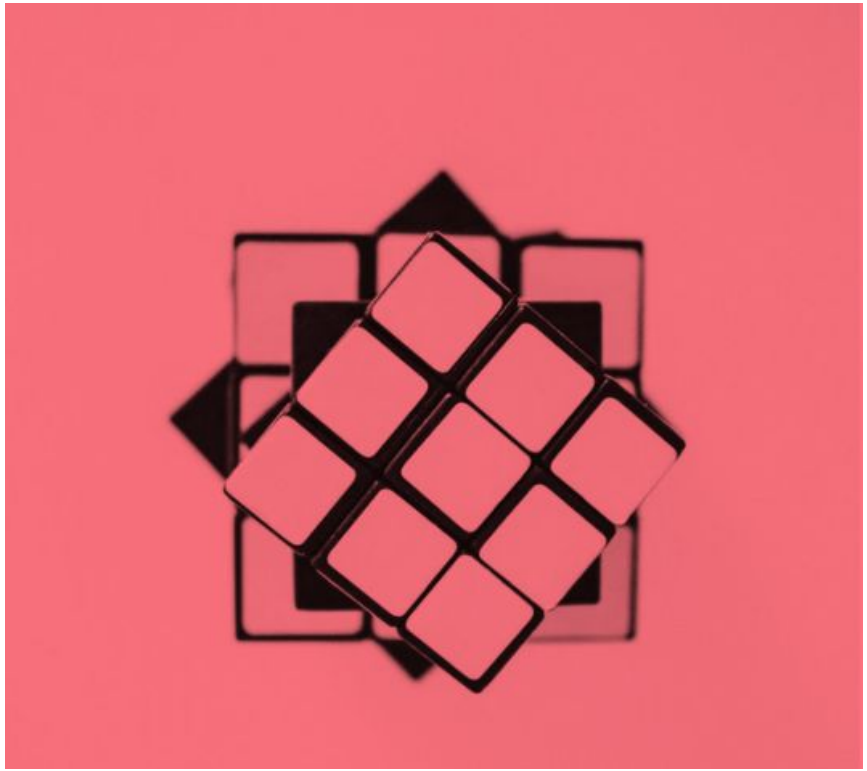
```
# df_plot = df_links.sort_values('weight', ascending=False).head(150)
df_plot = df_links[df_links['weight']>6]
df_plot.reset_index(inplace=True, drop=True)

G_plot = nx.Graph()

for link in tqdm(df_plot.index):
  G_plot.add_edge(df_plot.iloc[link]['from'],
                  df_plot.iloc[link]['to'],
                  weight=df_plot.iloc[link]['weight'])
```

```
100%|██████████| 156/156 [00:00<00:00, 2258.36it/s]
```

```
pos = nx.kamada_kawai_layout(G_plot)
nodes = G_plot.nodes()
fig, axs = plt.subplots(1, 1, figsize=(15,20))

el = nx.draw_networkx_edges(G_plot, pos, alpha=0.1, ax=axs)
nl = nx.draw_networkx_nodes(G_plot, pos, nodelist=nodes, node_color='#FF427b',
                            node_size=50, ax=axs)
ll = nx.draw_networkx_labels(G_plot, pos, font_size=10, font_family='sans-serif')
```



Results- Prototype



Key Challenges

- Removing duplicates from our data
- Lack of an industry standard graph evaluation technique
- Difficulty in Data scraping from reports
- Improving the accuracy of our entity extraction technique
- Size of data and time to compute

Next Steps

- Merging graph edge metrics/building multiplex graph
- Evaluation of community detection on test set of data points
- Fine-tune edge cutoffs and evaluate performance of graphs
- Formalize findings into a research format

Thanks

Any questions or comments?