Carnegie Mellon University

Practicum | Statistical Practice

Extracting graphical structures from mixed data sources

J.P.Morgan

Meet the Team

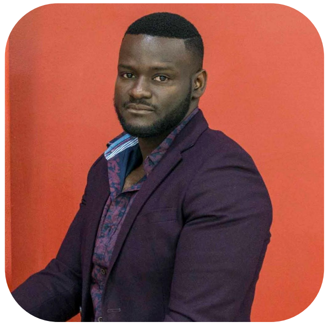
Supervisor

Aline Niyonsaba

MSc. Information Technology

Eric Ngabonzima

MSc. Information Technology



Ernest Kufuor Jr.

MSc. Information Technology

Ryan Harty

MSc. Statistical Practice

Dr. Moise Busogi

Ph.D. System Design and Control Engineering

Presentation Outline

1. The Why
2. Project Objectives
3. Solution Design
4. Apollo
    1. Demo
    2. Evaluation and Limitations
    3. Future Works

The Why?

J.P. Morgan 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 identify relationships between companies to help guide processes like investment strategy and fraud detection.

About JP Morgan AI Research

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.

J.P.Morgan

Project Objectives

Data Scraping

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

Building Knowledge Graphs

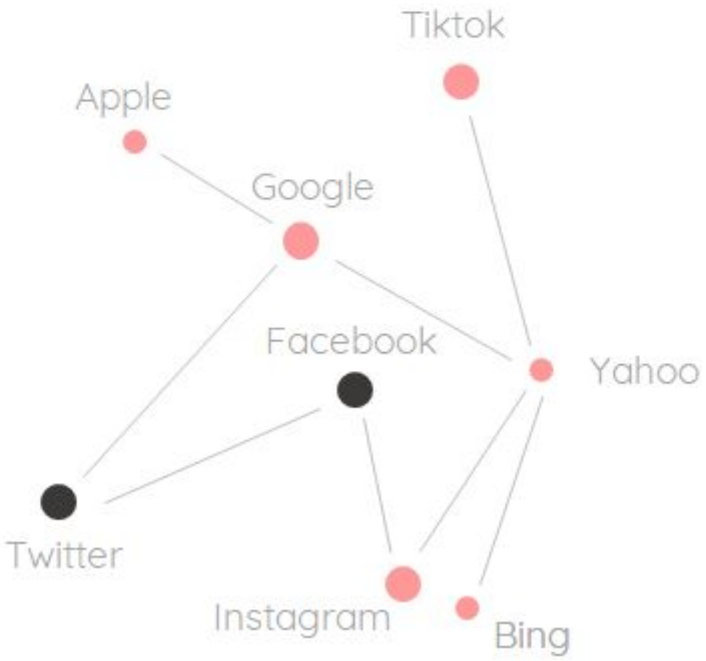
Pull entities from unstructured data and store into a knowledge graph.

Insights

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

Advance Knowledge

Publicize results to advance knowledge in this field.



Carnegie Mellon University

Solution Design

Users, Requirements and System Design

Users

1. Data scientist
2. AI researcher
3. Analyst
4. Developer

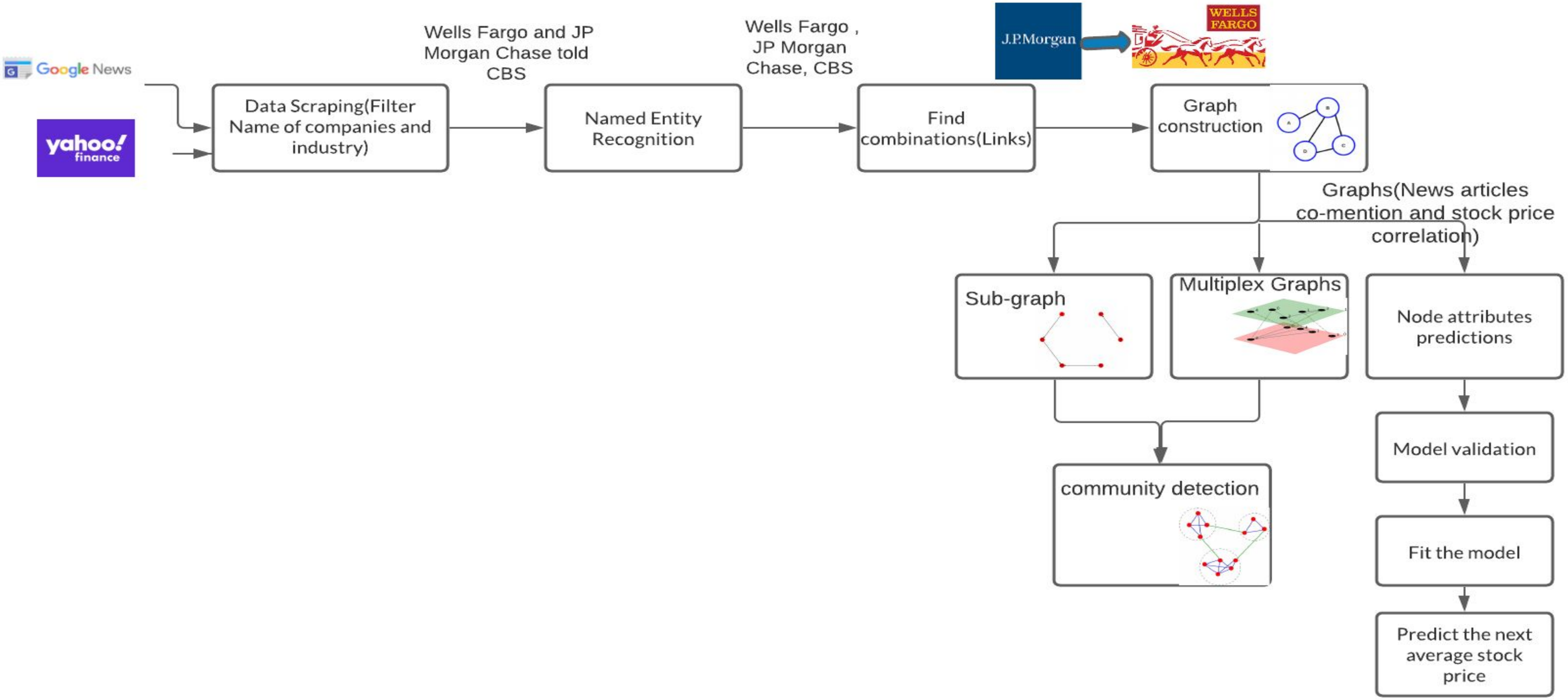
Functional Requirements

1. The system should be able to scrape data from Google News articles and Yahoo finance
2. The system should be able to extract entities from Google News articles' contents
3. The system should be able to build graphs
4. The system should be able to detect communities
5. The system should be able to predict node attributes
6. The system should be able to visualize sets of data.

Solution Design

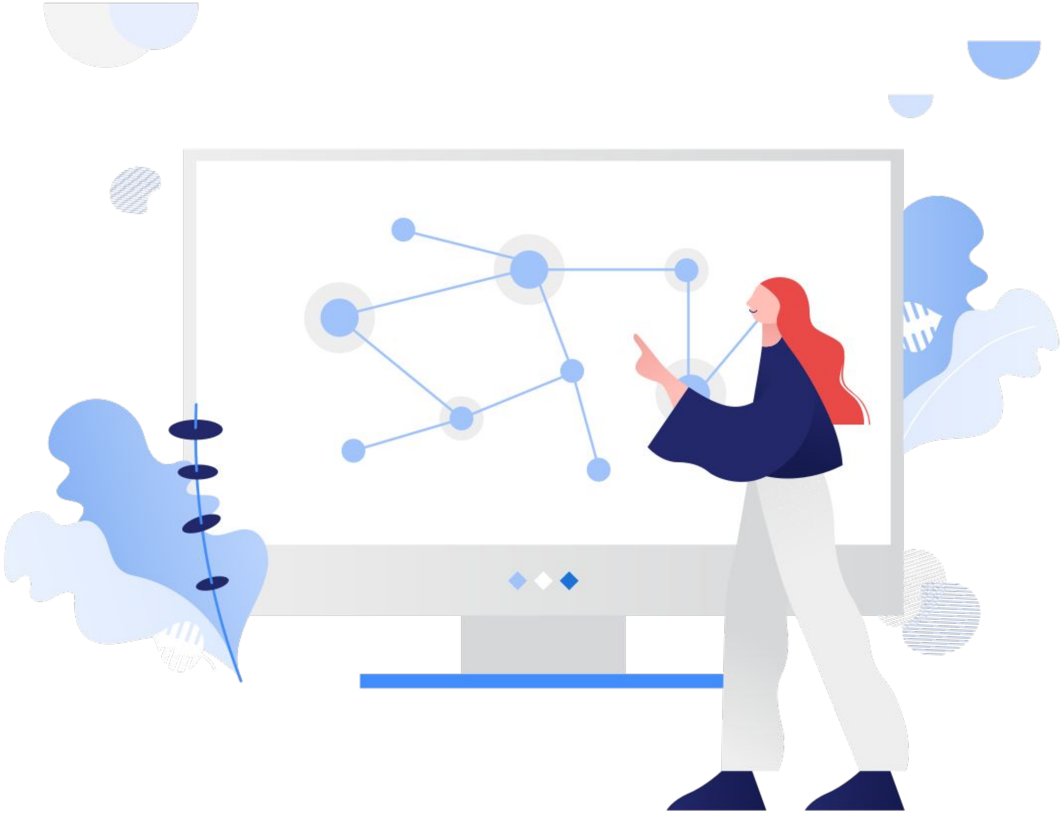
Block Diagram

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


Google News
yahoo! finance
Data Scraping(Filter Name of companies and industry)
Wells Fargo and JP Morgan Chase told CBS
Named Entity Recognition
Wells Fargo , JP Morgan Chase, CBS
Find combinations(Links)
J.P.Morgan
WELLS FARGO
Graph construction
Graphs(News articles co-mention and stock price correlation)
Sub-graph
Multiplex Graphs
Node attributes predictions
community detection
Model validation
Fit the model
Predict the next average stock price

Apollo

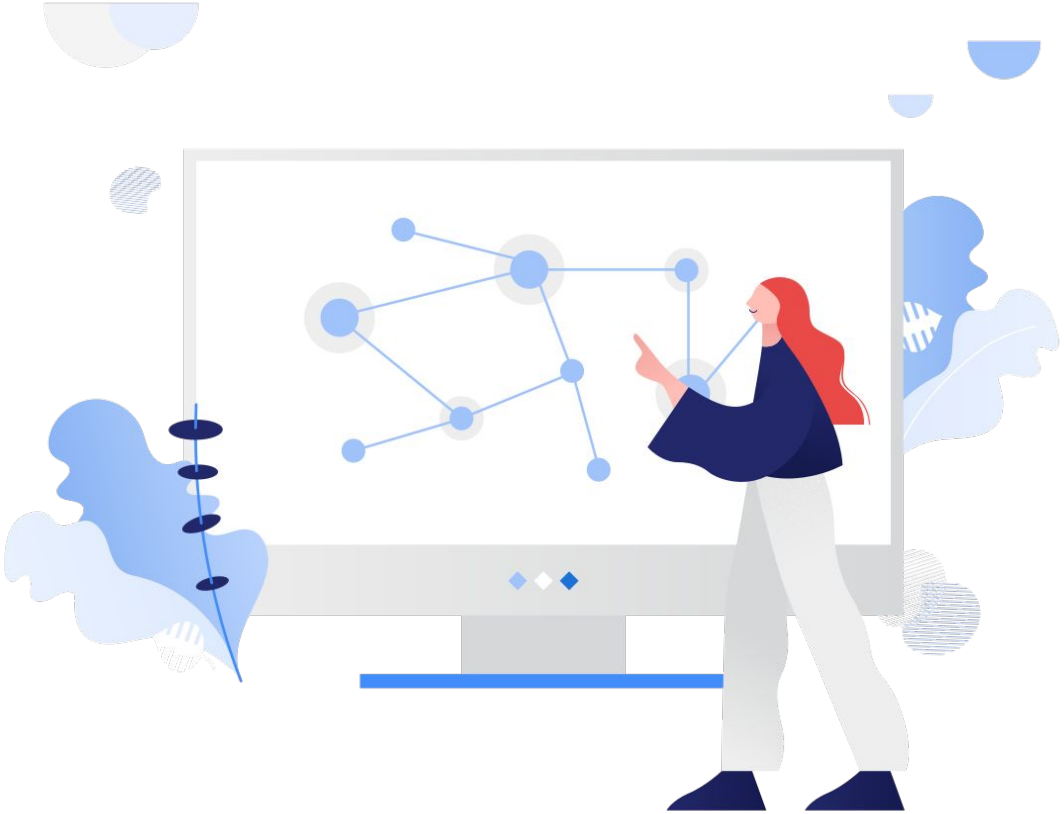
Python Package



Apollo

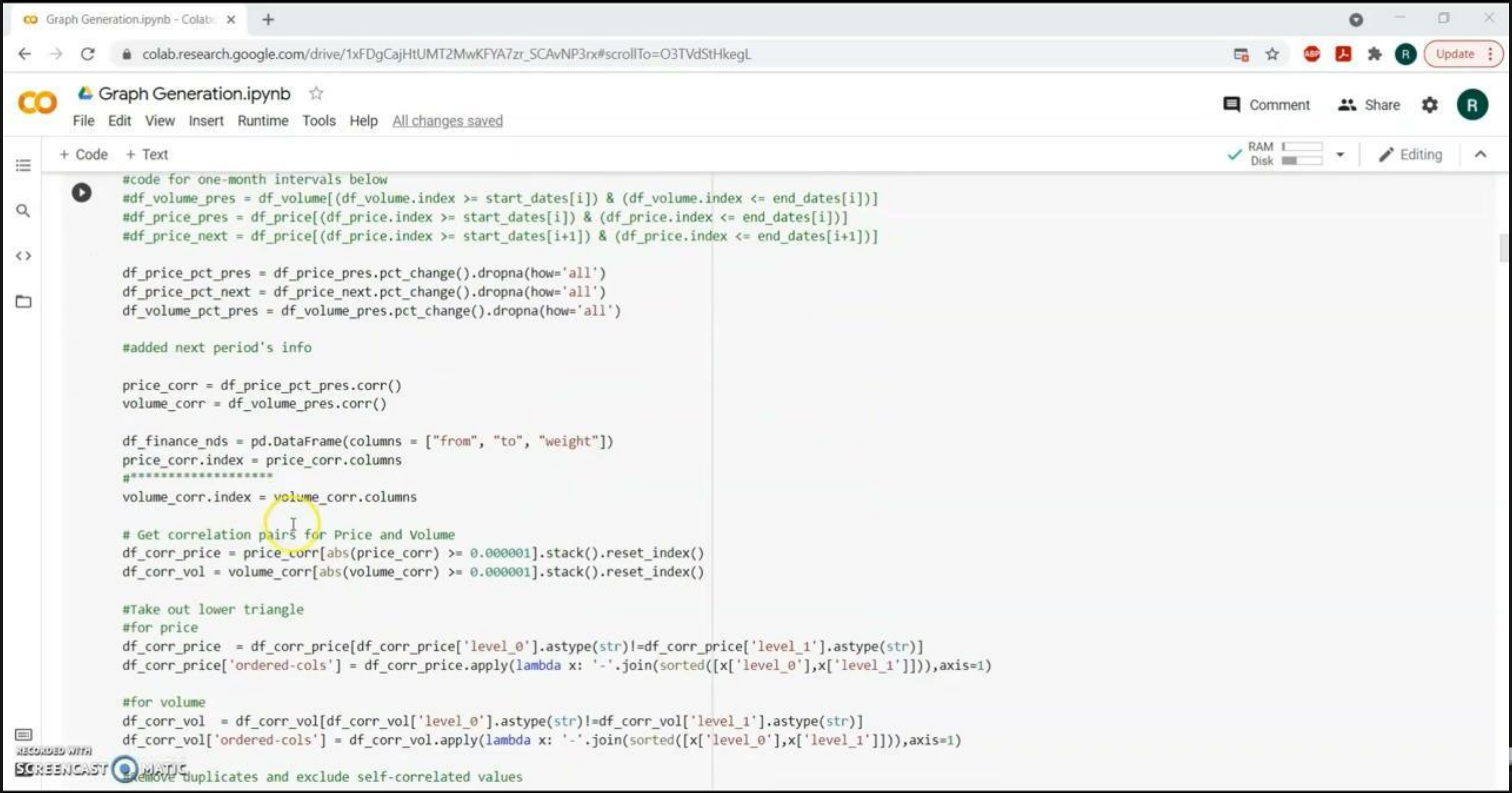
Python Package

1. News and Stocks Data Scraping
2. Graph Creation and Manipulation
3. Multiplex Graph Support
4. Graph Convolutional Network Support
5. Node Attribute Prediction



Apollo

Demo and Walkthrough


Graph Generation.ipynb - Colab
colab.research.google.com/drive/1xFDgCajHtUMT2MwKFYA7zr_SCAvNP3rx#scrollTo=O3TVdStHkegL
Update
Graph Generation.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing

```
#code for one-month intervals below
#df_volume_pres = df_volume[(df_volume.index >= start_dates[i]) & (df_volume.index <= end_dates[i])]
#df_price_pres = df_price[(df_price.index >= start_dates[i]) & (df_price.index <= end_dates[i])]
#df_price_next = df_price[(df_price.index >= start_dates[i+1]) & (df_price.index <= end_dates[i+1])]

df_price_pct_pres = df_price_pres.pct_change().dropna(how='all')
df_price_pct_next = df_price_next.pct_change().dropna(how='all')
df_volume_pct_pres = df_volume_pres.pct_change().dropna(how='all')

#added next period's info

price_corr = df_price_pct_pres.corr()
volume_corr = df_volume_pres.corr()

df_finance_nds = pd.DataFrame(columns = ["from", "to", "weight"])
price_corr.index = price_corr.columns
#********************
volume_corr.index = volume_corr.columns

# Get correlation pairs for Price and Volume
df_corr_price = price_corr[abs(price_corr) >= 0.000001].stack().reset_index()
df_corr_vol = volume_corr[abs(volume_corr) >= 0.000001].stack().reset_index()

#Take out lower triangle
#for price
df_corr_price  = df_corr_price[df_corr_price['level_0'].astype(str)!=df_corr_price['level_1'].astype(str)]
df_corr_price['ordered-cols'] = df_corr_price.apply(lambda x: '-'.join(sorted([x['level_0'],x['level_1']])),axis=1)

#for volume
df_corr_vol  = df_corr_vol[df_corr_vol['level_0'].astype(str)!=df_corr_vol['level_1'].astype(str)]
df_corr_vol['ordered-cols'] = df_corr_vol.apply(lambda x: '-'.join(sorted([x['level_0'],x['level_1']])),axis=1)

#Remove duplicates and exclude self-correlated values
```

Apollo

Evaluation and Limitations

Metric: Accuracy

For the experiment, we are using an evaluation set of 2018 – 2019 data for the node attribute prediction.

Training set: 2011 – 2017

Test set: 2018 -2019

Best Combined Graph Accuracy: 67.28%

Best Knowledge Graph Accuracy: 66.36%

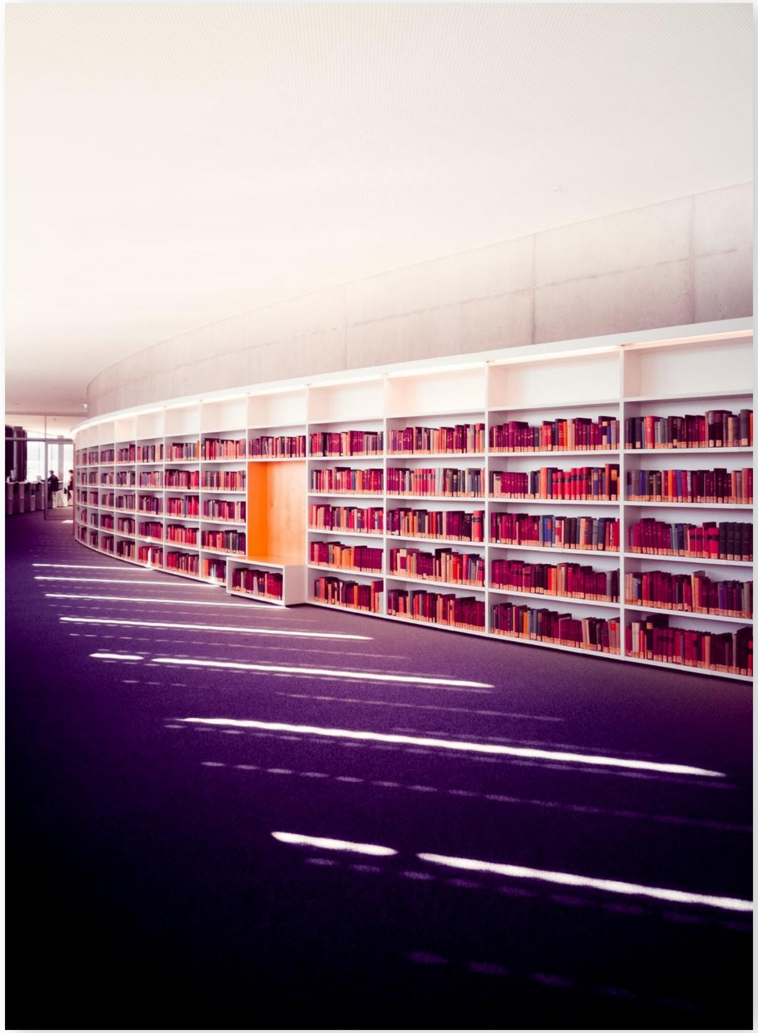
Limitations

1. Dataset size
2. Computing power

Apollo

Future Work

1. Tuning neural network parameters to build the best network
2. Tuning the graph parameters to build the best graph
3. Overcoming HTTP Timeouts to get a smoother news data pull from Google News
4. Analyzing multiple data sources to find which sources can provide a better combination
5. Analyzing multiple prediction tasks for the node attribute prediction



Thank You

All Questions and comments are welcome.