CS 6350
ASSIGNMENT _3_

Names of students in your group:

Kuei-Yu Tsai (kxt230002)

Number of free late days used: _0_

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

1. Finnhub News API
   https://finnhub.io/docs/api/news
2. JohnSnowLabs Spark-NLP:
   https://github.com/JohnSnowLabs/spark-nlp
3. spaCy
   https://spacy.io/models/en#en_core_web_sm
4. Apache Kafka
   https://hub.docker.com/r/apache/kafka
5. Elasticsearch and Kibana
   https://www.elastic.co/docs/deploy-manage/deploy/self-managed/install-kibana-with-docker
6. Amazon product co-purchasing network and ground-truth communities
   https://snap.stanford.edu/data/com-Amazon.html

1. Spark Streaming with Real Time Data and Kafka

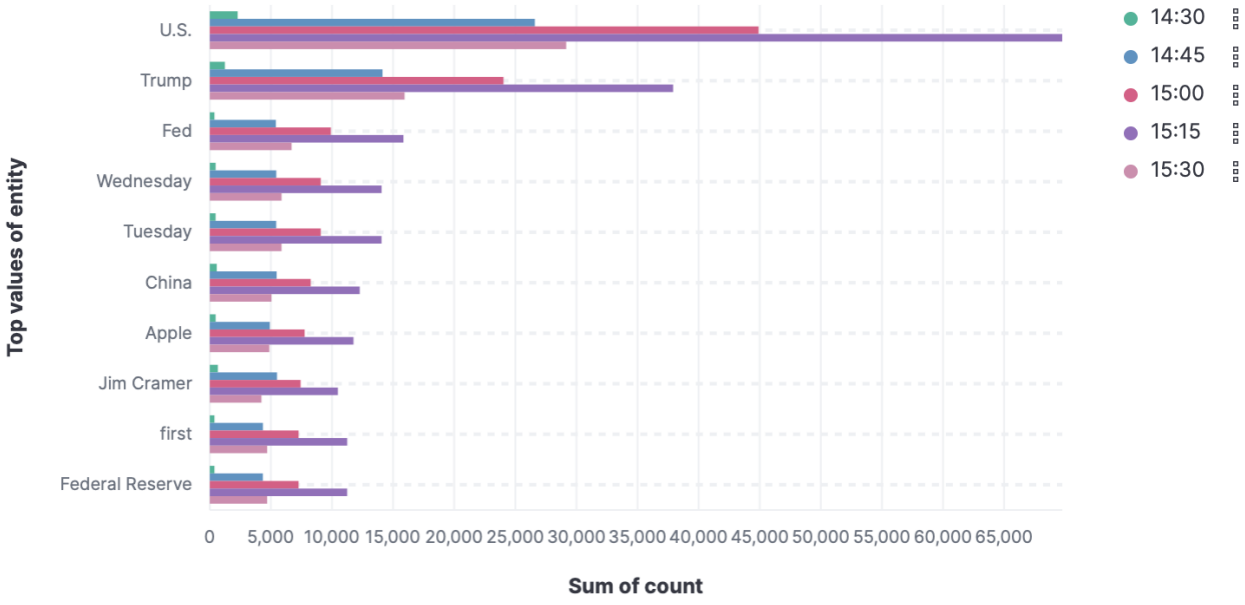
Data Source

The data source used in this project is the Finnhub News API, which provides real-time financial news headlines and summaries. We accessed general market news using their ` general_news('general', min_id=0)` endpoint via producer.py. News items were streamed to a Kafka topic (`topic1`) every 30 seconds. Each item includes a `headline`, `summary`, and `datetime` field.

Named Entity Recognition (NER) was performed using the `spaCy` NLP library with the `en_core_web_sm` model. Extracted entities were streamed into Kafka `topic2` via Apache Spark Structured Streaming and then ingested into Elasticsearch using Logstash. Finally, the entity frequency data was visualized with Kibana.

Results and Interpretation

The visualization shows the Top 10 entities detected in news articles at different time intervals:



Key Observations:
- "U.S." is the most dominant entity across all intervals, indicating that most financial news focused on U.S. policies, economy, or markets.
- "Trump", "Fed", and "Federal Reserve" consistently appear, likely due to ongoing economic or political developments during the period.
- Weekday-related entities like "Wednesday" and "Tuesday" are also frequent, suggesting temporal references in headlines.
- Notable figures like "Jim Cramer" and companies like "Apple" show moderate frequency, reflecting company-specific financial commentary or earnings news.

The consistent patterns over each interval indicate stable topic trends, with only slight variations in the frequency of entity mentions. This suggests that while new headlines arrive in real time, the focus of media coverage remains relatively consistent within a short temporal window.

Conclusion

This real-time NER pipeline effectively tracks the most discussed financial entities over time, allowing us to monitor trending topics, individuals, and organizations in the news. The modular architecture (Kafka + Spark + ELK) ensures scalability for larger news volumes or extended analysis periods.

2. Analyzing Social Networks using GraphX/GraphFrame

Output File with Results of Queries:

a. Top 5 nodes with highest outdegree

id	outDegree
21	62166
7	53550
6	45977
11	40290
1	26966

b. Top 5 nodes with highest indegree

id	inDegree
21	62166
7	53550
6	45977
11	40290
1	26966

c. Top 5 nodes with highest PageRank

id	PageRank
21	6384.880075830329
6	5652.756230220395
7	5362.062282094499
11	4306.846780175733
1	3595.660142415134

d. Top 5 connected components by number of nodes

Component ID	Node Count
1	306995
278	344
4203	253
448	131
9477	120

e. Top 5 nodes by triangle count

id	Triangle Count
21	238877
7	199434
11	112001
6	100261
38	99971

Summary and Insights

This analysis reveals several key characteristics of the social network graph:

Central Nodes:

Nodes 21, 7, 6, and 11 consistently rank highest across outdegree, indegree, PageRank, and triangle count, indicating they are central hubs with significant influence and strong local clustering.

PageRank & Triangle Count:

The PageRank scores correlate with high degree nodes, reinforcing their importance. Node 21 especially stands out, with the highest degree and triangle count, suggesting it is embedded in tightly-knit communities.

Community Structure:

The graph contains multiple disconnected components, with the largest component comprising over 300,000 nodes. This indicates the network is mostly connected but still has smaller isolated sub-networks.

Dense Clustering:

The triangle count analysis shows that key nodes participate in a vast number of closed triplets, implying the presence of community-like structures or cliques.

In summary, the results make sense and provide strong insights into both node-level importance and global structure of the network. These patterns are consistent with what we would expect from real-world large-scale social network data.