CS 6350

ASSIGNMENT \_\_\_\_\_3\_\_\_\_\_\_\_

Names of students in your group:

* Apoorva A Pise - AAP190023
* Tirth B Raval - TXR210000

Number of free late days used: \_\_\_\_\_1\_\_\_\_\_  
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.

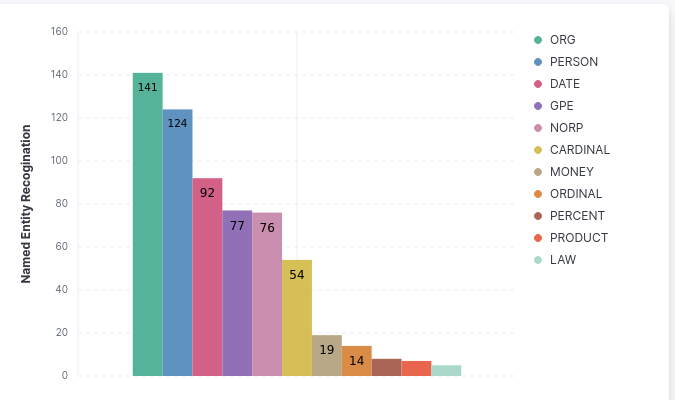
Part\_1:

Data Source:

* We utilized praw, a Python tool that is a wrapper for the Reddit API, to easily retrieve information from Reddit such as posts, comments, and user details. Programmers may easily extract named entities such as names of individuals, organizations, and locations from Reddit text data by integrating praw with spacy, a strong natural language processing software. This gathered data can then be processed to acquire insights into user sentiments, categorize articles and comments, and investigate discussions about certain themes or people.

Results:

Results of first 15 minutes of Reddit comments and we can see that the ORG tag has the greatest number of count.



Results at 30 minutes, here we can see that Reddit is producing the ORG tag the most

Chart, bar chart

Description automatically generated

Results at 60 minutes, here we can say that top most value is ORG and the least value is Time, early it was Money

Chart, bar chart

Description automatically generated

Results at 60 minutes, the top most value is still ORG and in the last iteration we got the highest value in ORG

Chart, bar chart

Description automatically generated

Part\_2: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/7856442364476838/2415894614121183/3680171464303125/latest.html>

* Input Dataset Path: <https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html>
* Output Dataset Path: <https://utdallas.box.com/s/ym4xuufslzxhsp1lpsvenb8b82cjok39>

**Result of queries:**

a. Find the top 5 nodes with the highest outdegree and the count of the number of outgoing edges:

Table

Description automatically generated

b. Find the top 5 nodes with the highest indegree and the count of the number of incoming edges

Table

Description automatically generated

c. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values

Table

Description automatically generated

d. Find the top 5 components with the largest number of nodes.

Table

Description automatically generated

e. Find top 5 vertices with the largest triangle count.

Table

Description automatically generated

**Summary:**

* Outdegree: This analysis can reveal which nodes in the Bitcoin OTC web of trust network have the highest number of outgoing links to other nodes, indicating their level of engagement with other nodes in the network and potentially indicating influential players in the Bitcoin OTC ecosystem.
* Indegree: This analysis can provide insights into which nodes receive the most incoming links from other nodes, indicating their significance or influence within the Bitcoin OTC web of trust network.
* PageRank: PageRank measures the importance of a node in a graph based on its connections with other nodes. This analysis can help identify the most important or influential nodes in the Bitcoin OTC web of trust network.
* Connected Components: Connected components are groups of nodes that are connected to each other either directly or indirectly. This analysis can help identify clusters or communities within the Bitcoin OTC web of trust network.
* Triangle Count: Triangles are a measure of clustering or community structure in a graph. This analysis can provide insights into which nodes have the densest interconnections, indicating strong community engagement or interactions in the Bitcoin OTC web of trust network.

**README:**

The code is run using data bricks notebook link provided above.

The path to the dataset is used in the arguments which directly load the dataset into the notebook. To run the code, run all the cells in the notebook.