Pagerank Cryptocurrency Price Prediction

Harsha Somisetty

May 3rd, 2021

Introduction and Problem

The past year has seen a huge rise in retail investings in traditional equity markets, and emerging Cryptocurrency markets. This trend has been amplified by social media, specifically through the abundance of information available through services that allow for rapid sharing of information like Twitter, Reddit, Telegram, and Discord. However, there is too much information and too little time to sort through to find good investment opportunities before others. Traditionally, people have used streaming algorithms combined with sentiment analysis on a general stream of information (tweets, chatlogs) as research, however, this method lacks the incorporate of trust. Specifically, bots can be used to manipulate sentiment and mentions of a particular asset, or random people with no real investment experience might cloud signals from a system. Accordingly, this project aims to address these problems through the use of data mining techniques.

Solution and Implementation

To narrow the scope of the project, data will be only collected from Twitter, and would only be regarding cryptocurrencies (coins) traded on the Binance Exchange. Again, the goal of the project is to aggregate mentions of assets from trusted users, and to use that data to automatically find new potential trades. To do this, we create a Neo4j graph database of people information, specifically the people who they follow and are followed by, as well as what coins they mention.

The graph would therefore consist of 300 trusted users, who those users follow within the 300 users, and what coins each user has mentioned in the past.

Since coins mentions can happen any number of days in the past any number of times, we devised a method to calculate the strength of a coin mention as follows. For a certain tweet mentioning a coin, we find the difference in days between the tweet date and the current time of the pagerank calculation, denoted as x. We then normalize this difference by considering $e^{-.08x}$, which allows us to weight a tweet 1 month ago about BTC as less relavant than a tweet about BTC 1 day ago. Then for all of a user's recent 2500 tweets, we add all the normalized values, and add them up. This final value is the strength of connection between a person and a coin. This process is repeated for every single person, and all the edges between people and mentions are generated.

After loading this data, pagerank is conducted on this graph of users, who they follow, and what coins they mention, and scores of the most relavant coins are calculated. But since this calculation doesn't take into account the sentiment of each user about a coin, the pagerank is representative of the social volatility of a coin.

Thus, to identify assets before they become mentioned more and more, we consider the pagerank scores of the coins over time. Specifically, we will consider tweets in a window of 90 days, calculate

The specific steps of the algorithm are as follows:

Data Collection

1. Collect list of people on Twitter
2. Download everyone each person follows
3. Download all tweets of each person

Data Processing

Graph Database Creation

1. For each user in the list, gather all the intragroup follows, edge weight is reciprocal of number of people they follow
2. For each user's tweet history, select window of 30 days and extract coin mentions as explained. Each coin mention and weight becomes an edge
3. Combine these follow and mention edges, and user and coin nodes into a single graph, and run weighted pagerank. Collect scores of coin nodes

Evaluation

Conclusion and Future improvements