Data Visualization Project

MASTER'S DEGREE PROGRAM IN

DATA SCIENCE AND ADVANCED ANALYTICS

EXPLORING CRYPTO REDDIT

Group 10

Poles Natalia, number: 20210675

Jannuzzi Marcelo, number: 20210674

Parenti Alberto, number: 20211304

04/2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Links

Links to the other final project deliverables:

- GitHub repository
- Dash app deployed on Heroku

Dataset description

We chose to work with a dataset containing data from Reddit, a social media platform that allows users to post and comment on content. The different communities that exist on Reddit are organized into groups called *subreddits*. A subreddit is formed around a specific subject, which can be general (such as science or movies) or very specific (such as bonsai trees or helping to identify typefaces).

The dataset we selected contains almost 5 years of data from Reddit, specifically, it contains data from Reddit posts which link to other subreddits, including the source and target subreddits, the post's timestamp, sentiment score and many other characteristics of the post. This subreddit-to-subreddit hyperlink network allows us to explore how different subreddits are connected, as well as the different characteristics of each subreddit.

Visualization and interaction choices

We wanted to work on a project related to graph networks, and after researching different datasets available with such information, we were attracted to the fact that this one contained a lot of data which was collected from the interactions of real users.

After combining the 2 main datasets available (one related to the titles of the Reddit posts and the other related to their bodies), we were left with around 900 thousand lines of data on approximately 350 thousand subreddits. Because of this large volume of data and due to the fact that most of these subreddits are completely unrelated, we chose to work with only a small subset of subreddits related to cryptocurrencies (which is a topic of personal interest to some members of the group). Also, from personal experience we found that plotting a graph with even a few thousand nodes to be infeasible due to both computer memory and interpretability issues, so limiting our project to this particular subject also allowed us to tell a better story with the data.

As mentioned above, the central information contained in this dataset is the network of subreddits connected to each other through hyperlinks, so we made this graph the main aspect of our dashboard. As explained below in the *Technical Aspects* section, we implemented this graph using the Dash Cytoscape library, which allowed us to make the network very interactive. The user can zoom in and out, pan, and select a node in the graph to bring it into focus and into the forefront of the visualization, as well as get a bit more information on its connections and to make it clearer to see what its direct neighbors are.

We have also allowed for some interactivity in the other panel placed beside the network graph: the user can select one or more subreddits from the dropdown list, as well as the sentiment of these subreddits' posts, and see extra information about these filtered posts.

Reading the visualization

The first graph shows the connections between the subreddits. The node size is directly proportional to the amount of posts in that particular subreddit. On the other hand, the edges connecting two nodes are thicker as the number of connections between the two subreddits increases. By clicking on a particular node, its edges get colored with a diverging 2-color scheme encoding the average sentiment of the connection: more red means a more negative sentiment, and greener means a more positive one.

Top chart	
Data Items	Crypto subreddits
Data Attributes	Number of posts, connections and average sentiment
Visual Marks	Points for each subreddit and lines for their connections
Visual Channel	Point size, line width and line color
Data Encoding	Number of posts - point size; average sentiment between subreddits - line color; number of connections between subreddits - line width

Using the dropdown list on the top side of the second panel it is possible to select multiple crypto subreddits and the sentiment the viewer looks for: *neutral or positive*, and *negative*. Thus, the information about the selected posts appears on the table below showing the amount of words, unique works and stop words, the fraction of stopwords, the automated readability index and finally the compound sentiment calculated by VADER [1]. These node attributes were calculated by researchers at Stanford University [2, 3].

The final panel on the bottom of the dashboard shows a bar plot with the sentiment distribution in all crypto subreddits.

Bottom chart	
Data Items	Crypto subreddits
Data Attributes	Type of sentiment, count
Visual Mark	Lines (or bars)
Visual Channel	Bar size for the quantitative attribute and position on the y axis for the categorical attribute
Data Encoding	Type of sentiment - y axis; Number of posts - x axis

Technical aspects

Because the dataset we chose to work with was very large, we couldn't upload it to our GitHub repository, so we created a version of it which we already filtered and preprocessed. The code in `filter_dataset.ipynb` creates this filtered dataset in the root directory of our repository, titled `filtered_and_treated_dataset.csv`.

To create the graph network from the original dataset, we used the excellent NetworkX Python library, and to visualize the graph we relied heavily on the Dash Cytoscape library, which is a wrapper around Cytoscape.js, a library used to display and manipulate rich, interactive graphs.

Cytoscape proved to be a difficult library to use, and required us to do a lot of data wrangling in order to accomplish what we wanted, but the general workflow was:

1. Create the NetworkX graph from a Pandas DataFrame.
2. Iterate through the NetworkX graph nodes and edges to include attributes we wanted to use as channels in our visualization.
3. Create a Cytoscape graph from the NetworkX graph
4. Add the node and edge attributes to the Cytoscape graph since, for some reason, the convenience function to convert a NetworkX graph to a Cytoscape one (`nx.cytoscape_data()`) doesn't do this automatically.
5. Write callbacks to apply different stylesheets to the graph as the user clicks on different nodes.

Discussion

Getting a sense of how groups of entities are connected is very hard without a visualization of the network, which is why graphs are so important for this. The network we have plotted allows us to have a good sense of the overall characteristics of cryptocurrency-related subreddits, including: the main subreddits in terms of volume of posts, the subreddits which link the most to each other and the sentiments expressed between these subreddits.

However, there are many additional features we wanted to add to our graph, but which proved challenging to make work adequately. We would have liked to:

- Give the user the option to reset the graph to its initial (unhighlighted) state, after selecting a particular node to focus on.
- Make the nodes not overlap each other. We came up with a workaround for this though, which was to bring the highlighted nodes to the foreground once they were selected.
- Connect the graph network to the dropdown filter to allow the user to only see the selected subreddits in the graph view.
- Add the graph legend as a native Plotly element, instead of as an image.

- Allow the user to choose what they wanted to see in the different marks. For example, instead of only encoding the number of posts as the size of the nodes, to allow the user to encode something else onto this channel, such as the average length of the posts in that subreddit.
- Add information regarding the direction of the connections. We chose to summarize the information contained in the links between subreddits as an undirected graph, but we did have information on the direction of the links (the dataset provided information on which were the source and target subreddits).
- Display an alternate visualization for the graph using a matrix view.

Furthermore, it would also be interesting to visualize other categories of subreddit in this dashboard, aside from only crypto-related subreddits. That way we could see how different arts or sports-related subreddits are connected, or even how different crypto-related subreddits are connected to political subreddits, therefore uncovering the political inclinations of each group.

References

[1] C.J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI, jan 2015.

[2] S. Kumar, X. Zhang, J. Leskovec. In *Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks*. ACM SIGKDD 2019.

[3] S. Kumar, W.L. Hamilton, J. Leskovec, D. Jurafsky. In *Community Interaction and Conflict on the Web*. World Wide Web Conference 2018.