Report on On-Demand Data Analytics Projects: Community Voting on 7 Most Sought-After Tools or Datasets Projects

1. Introduction about milestone 3

This milestone 3 report details the progress and outcomes of an on-demand data analytics project, focusing on community voting to identify the 7 most sought-after tools or datasets. The key objective for this milestone is to identify the 7 most voted issues and findings, along with the data analytics tools and datasets, that require urgent attention to prevent gaming and abuse across Catalyst fundings.

2. Community Voting Process

The community was engaged to vote on the most critical tools or datasets required to support ongoing and future fund data analytics projects. This voting process was designed to ensure that the selected tools/datasets align with the needs and priorities of the community. The outcome identified the top 7 most sought-after tools or datasets, which will be crucial for the success of future projects within the ecosystem.

The survey was conducted based on *14 potential directions for deep-dive analysis, including GitHub data reports and community-submitted issues* [ref(1)]. Over a week of voting, the survey received a strong response. Here are the top 8 dataset projects the community voted to see, listed in the same order as in the *Typeform survey & survey result* [ref (2)]. Due to tied scores, we are including 8 projects instead of the originally planned 7.

1.**Proposal Keyword Occurrences vs Funding Success**

- Approach: Use Natural Language Processing (NLP) with Pandas and Scikit-learn to analyze keyword frequency in proposals. Compare keyword occurrences with funding success rates using statistical correlations or machine learning models.
- Tools: Pandas, Scikit-learn, NLTK or SpaCy for NLP

2.**Boxplot of Proposal Assessor Scores vs Funding Success**

- Approach**:** Aggregate assessor scores for each proposal and categorize them by Funding success. Use Matplotlib or Seaborn to create boxplots that visualize the distribution of scores across successful and unsuccessful proposals.
- Tools**:** Pandas, Matplotlib, Seaborn.

3.**Votes Required Evolution: Funding Request by Fund and Status**

- Approach: Track the evolution of votes required for funding requests across different funds and statuses. Use Pandas for data aggregation and Matplotlib to create time series or bar charts showing the trend over time.
- Tools: Pandas, Matplotlib

4.**YES votes Required for Funding (by categories) across Funds**

- **Approach:** Calculate the number of YES votes required for funding across different categories and funds. Visualize the data using stacked bar charts or line plots to compare categories.

- **Tools:** Pandas, Matplotlib, Seaborn.

5. **Networks of Groups with large numbers of Proposal Submissions per Fund per Entity**

- **Approach:** Build a network graph to represent the relationships between entities and the number of proposals they submit per fund. Use NetworkX for constructing and visualizing the network.

- **Tools:** Pandas, NetworkX, Matplotlib.

6. **Data-Processing of Milestone Module: Reading as JSON, tokenization, insights**

Please refer to the telegram chat by a community member (Yuta)

https://docs.google.com/document/d/1DDb4PIUGSHLaAqie0Jt_PZ3kCfjl56my/edit?usp=sharing&ouid=114192021221088455069&rtpof=true&sd=true

and GitHub issue logged by another community member (saigonbitmaster)

https://github.com/Sapient-Predictive-Analytics/Data-Driven_Catalyst/issues/5

For all issues logged by the community on GitHub, please refer to [ref(3)]

- **Approach:** Read and process milestone data from JSON files. Tokenize text data to extract insights, such as milestone completion trends or common challenges. Use Pandas for data manipulation and NLTK or SpaCy for tokenization.

- **Tools:** Pandas, NLTK or SpaCy, JSON module.

7. **Rotational Breaks for Big Winners: The Impact of Past Success on Current Funding**

- Approach: Analyze the relationship between a team's past funding successes and their chances of receiving current funding. Use Pandas to track historical funding data, and apply statistical analysis or logistic regression to assess the impact of past wins on new funding opportunities.

- Tools: Pandas, Scikit-learn (for regression analysis), Matplotlib or Seaborn (for visualizations)

8. Scamming Catalyst: Cloned and Copied Ideas/Proposals Submitted Last Minute

Please refer to GitHub issue that was logged by a community member (Udai Solanki)

https://github.com/Sapient-Predictive-Analytics/Data-Driven_Catalyst/issues/8

- **Approach:** Use Natural Language Processing (NLP) techniques to detect similarities between proposals, especially those submitted close to deadlines. Compare these proposals with earlier ones to identify potential cloning or plagiarism.
- **Tools:** Pandas, NLTK or SpaCy (for NLP and similarity detection), FuzzyWuzzy (for string matching), Matplotlib (for reporting findings).

3. Feedback and Documentation

Following the voting process, feedback was solicited from the community to ensure that the selected tools and datasets met the diverse needs of stakeholders. This feedback was thoroughly documented and has been made publicly available on GitHub and GitBook. The documentation provides transparency and allows for continuous community engagement as the project evolves.

The community voting result and the screenshot of the survey form is in excel spreadsheet in this google drive:
https://docs.google.com/spreadsheets/d/180HgmAnyaoC9We4iSUSK1WcEAVYN9q7H/edit?usp=sharing&ouid=114192021221088455069&rtpof=true&sd=true

4. Integration with Catalyst System

Close liaison with the Catalyst Team was maintained throughout the project to ensure that the chosen tools and datasets could be seamlessly integrated into the ongoing liquid democracy implementation. This collaboration also involved aligning the process and licenses with the requirements of the Catalyst Voices initiative. The outcome is a flexible, well-documented approach that supports the broader goals of liquid democracy and ensures that future projects can build on a solid foundation.

Please refer to our communication with Catalyst Team on 13 Aug 2024

https://docs.google.com/document/d/1DDb4PIUGSHLaAqie0Jt_PZ3kCfjl56my/edit

We have also communicated with Catalyst Voice Team member but the response has been slow as the Catalyst Voice team has experienced the delay on their side (as per this telegram evidence and the change request of Catalyst Voice Team which we mentioned earlier)

https://drive.google.com/file/d/1OVO16cPh2PDpYDbriobLrLHbpzOxSMr9/view?usp=sharing

We have another upcoming meeting with Catalyst team to discuss further on our project as per this office hour link

https://drive.google.com/file/d/142zYT7wPWDoUGi5JphGNKfqvmaHF8f4u/view?usp=sharing

5. Next milestone

This milestone 3 has successfully identified the community's top priorities in terms of tools and datasets, ensured their alignment with liquid democracy efforts, and formalized the process through public documentation. Moving forward, these results will guide future developments and support ongoing innovation for the Catalyst Projects.

We already conducted an appointment to liaise with Catalyst Team and Catalyst voice on the milestone review process on August 13th. Please find *Google Meet screenshots for our liaison with Catalyst team* in this milestone [ref (4)]. Please note that the meeting was not recorded as per the request of the participant. However, the discussion was highly productive, and we provided Catalyst team with feedback on the milestone module process. We will also meet the Catalyst Team again later this month on Aug 30 to discuss the progress since our last talk and keep milestone reviewers and the community informed in our next milestone 4 submission, where we will already produce a comprehensive report on community-tasked selected projects. Thanks for your time for this milestone.

References

1. GitHub Origin of the 14 possible directions for deep-dive analysis: Github data reports & Community submitted Issues: https://github.com/Sapient-Predictive-Analytics/Data-Driven_Catalyst/blob/main/Funds/examples.md

2. Survey & Survey result: https://docs.google.com/spreadsheets/d/180HgmAnyaoC9We4iSUSK1WcEAVYN9q7H/edit?usp=sharing&ouid=114192021221088455069&rtpof=true&sd=true

3. GitHub Issues: https://github.com/Sapient-Predictive-Analytics/Data-Driven_Catalyst/issues

4. Evidence of Sapient Team communication with Catalyst team, Catalyst Voice and the next scheduled call with Catalyst Team

   https://docs.google.com/document/d/1DDb4PIUGSHLaAqie0Jt_PZ3kCfjl56my/edit?usp=sharing&ouid=114192021221088455069&rtpof=true&sd=true