SE 491 WEEKLY REPORT 6
10/17/2024 - 10/24/2024

Group number: sdmay25-30

Project title: *Explainable AI for source code applications*

Client &/Advisor: *Arushi Sharma*

Team Members:
Manjul Balayar
Kellan Bouwman
Sam Frost
Akhilesh Nevatia
Ethan Rogers

Weekly Rotation
Manjul Balayar
Kellan Bouwman
Sam Frost
Ethan Rogers
Akhilesh Nevatia

Weekly Summary

This week, the team made significant progress across various areas of the project. Alignment refactoring was completed and integrated on Colab, with initial efforts to incorporate lexical and contextual criteria into the pipeline. Testing and restructuring efforts focused on pip installability and transitioning the Colab workflow to a command-line interface format. The first round of unit tests for NeuroXCode_Test was completed. Clustering was fully refactored, with added test cases and code for demonstrating clustering functionality on Colab. Additionally, the process activations were entirely refactored, with the directory structure for data extraction implemented, example test code added to the local test bench, and test code transferred to Colab.

- **Past week accomplishments**

  Manjul Balayar:
  - Completed alignment refactoring and made sure it's imported on colab
  - Went over lexical and contextual criterias to add them into our pipeline
  - Tried to make an integrated pipeline on colab, somewhat worked but alignment was manual an score was 62%

  Kellan Bouwman:
  - Worked on testing for pip installability
  - Worked on setting up and structuring an example workflow with colab
  - Testing Colab Notebook
  - Restructuring Colab into a CLI accessible workflow

  Sam Frost:
  - Finish first round of unit tests for NeurXCode_Test
  - Research IEEE/ISO standards that could apply to our project

  Akhilesh Nevatia:
  - Refactored Clustering Completely
  - Tested functionality of Clustering added Test cases in the NeuroXCode_Test Folder
  - Added code to the Colab for Clustering, showing how someone can use the clustering part of the library

  Ethan Rogers:
  - Completely refactored process activations
  - Implemented desired structure of directories for data extraction
  - Added example test code to local test bench
  - Moved test code to Colab

- **Pending issues**

  Manjul Balayar: N/A
  Kellan Bouwman: N/A
  Sam Frost: N/A
  Akhilesh Nevatia: N/A
  Ethan Rogers: N/A

Individual Time Contributions

Name	Hours This Week	Total Hours
Manjul Balayar	6	37
Kellan Bouwman	6	44
Sam Frost	7	37
Akhilesh Nevatia	12	40
Ethan Rogers	7	38

- **Plans for the upcoming week**

Manjul Balayar:
- Finish up integrated pipeline before adding newer stuff
- Work on implementing lexical and contextual into evaluation
- Look into newer activation requirements

Kellan Bouwman:
- Design for next Sprint
- Implement Feedback for main methods / pipelines
- Work on improvements to colab (from feedback)

Sam Frost:
- Take feedback from initial pipeline demo and adjust accordingly
- Design for next sprint based on meeting
- Look into Hydra config, prepare to make a mock up once I understand it better

Akhilesh Nevatia:
- Help build Design for next Sprint
- Build an Integrated and functional pipeline on the colab
- Integrate Clustering with the Real Processed Activations with Ethan

Ethan Rogers:
- Design for next sprint
- Make process activations more user friendly and cohesive
- Decrease necessary function calls through multistep functions

Summary of weekly advisor meeting

We settled on a two-week sprint cycle, dedicating the first week to design and the second to implementation, with the past week focused on refactoring. Our next steps include implementing BIRCH, Hyperbolic (visualized using hypertree and agglomerative methods), Lexical, and Contextual criteria for clustering evaluation. The goal is to develop an integrated pipeline that matches the format used in CodeConceptNet, allowing us to compare different clustering algorithms and their ability to capture hierarchical structures. While standard clustering evaluation metrics have shown low scores, alignment remains good, suggesting the need to explore additional metrics. Given the limitations of current metrics, we will prioritize visualizing clustering results, as these insights may offer a more meaningful understanding of the data's structure.