(Joseph) Cole Ramos
07400
2 March 2022

**Weeks 5 to 7 - Milestone Report**

**Dates:** 16 February 2022 to 2 March 2022

Progress:

- Two main goals: to get the base language model running and evaluate its performance, and to get CodeQL running locally to find uses of undeclared variables in Python or Java
- The Language Model
    - Working with weights from my mentor's repository: https://github.com/VHellendoorn/Code-LMs
    - Using GPT-NeoX underneath as the toolkit for the model: https://github.com/EleutherAI/gpt-neox
    - Reading documentation for and configuring the model to run locally
    - Not entirely successful yet: it may be a waste of time to get it running locally, as the requirements for the large model may be beyond what I can do locally: I may need to discuss with my mentor more about running this on a server or in the cloud
- CodeQL
    - CodeQL is the best candidate for the annotator
        - Seems to be the rising industry standard in code analysis, and is quite flexible
    - Got it installed and running locally
    - Read documentation to understand more of the capabilities and how to use CodeQL
    - The inbuilt library queries for undeclared variables appear to only be for C/ C++: I can probably find a custom query to do this for Python or Java
    - Ran some local tests to see how the output of CodeQL worked for other commands: figuring out how to turn this output into an annotation will be one of the main functions I'll need to implement
    - The final CodeQL annotations likely won't use the UndeclaredVariableAccess query, but will rather use something similar (or possibly something from the control flow or data flow packages) to create backreferences from where a variable is being used to where it was declared