CLUSTERING PYTHON PROGRAMS

Aksel Õim, Ken Böckler

INTRODUCTION

The aim of the project was to find similar programs among Python programs submitted by students and cluster them. The goal was to detect collaboration and plagiarism among the students and improve automatic testing.

DATA

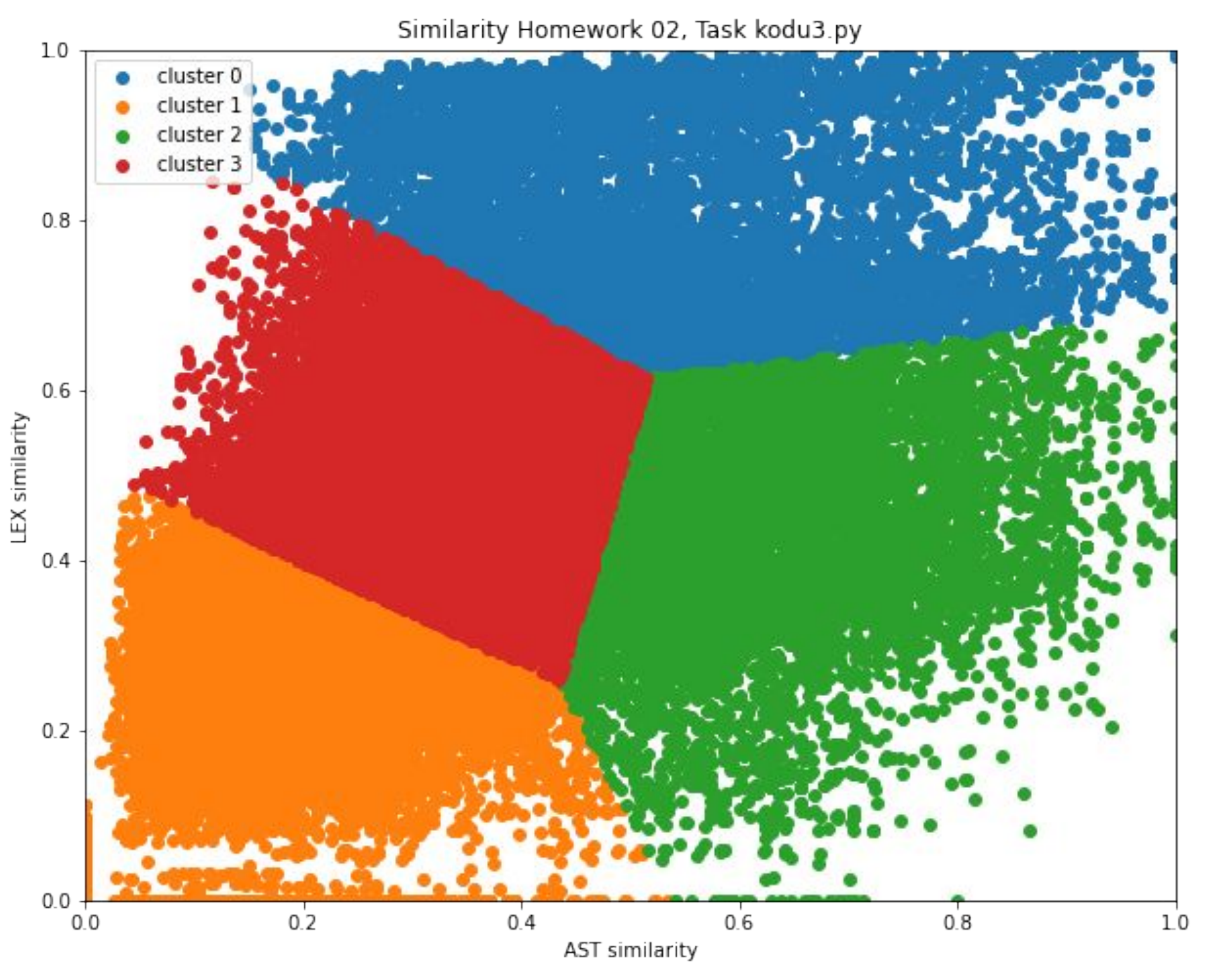
The data we used was submitted by students in the introductory Computer Programming course. All Together the data consisted of 13 homeworks and 33 tasks. Each homework had around 1 to 4 tasks.
We cleaned the data by removing every submission, except the final solution and then removed comments and empty lines from the final solution.

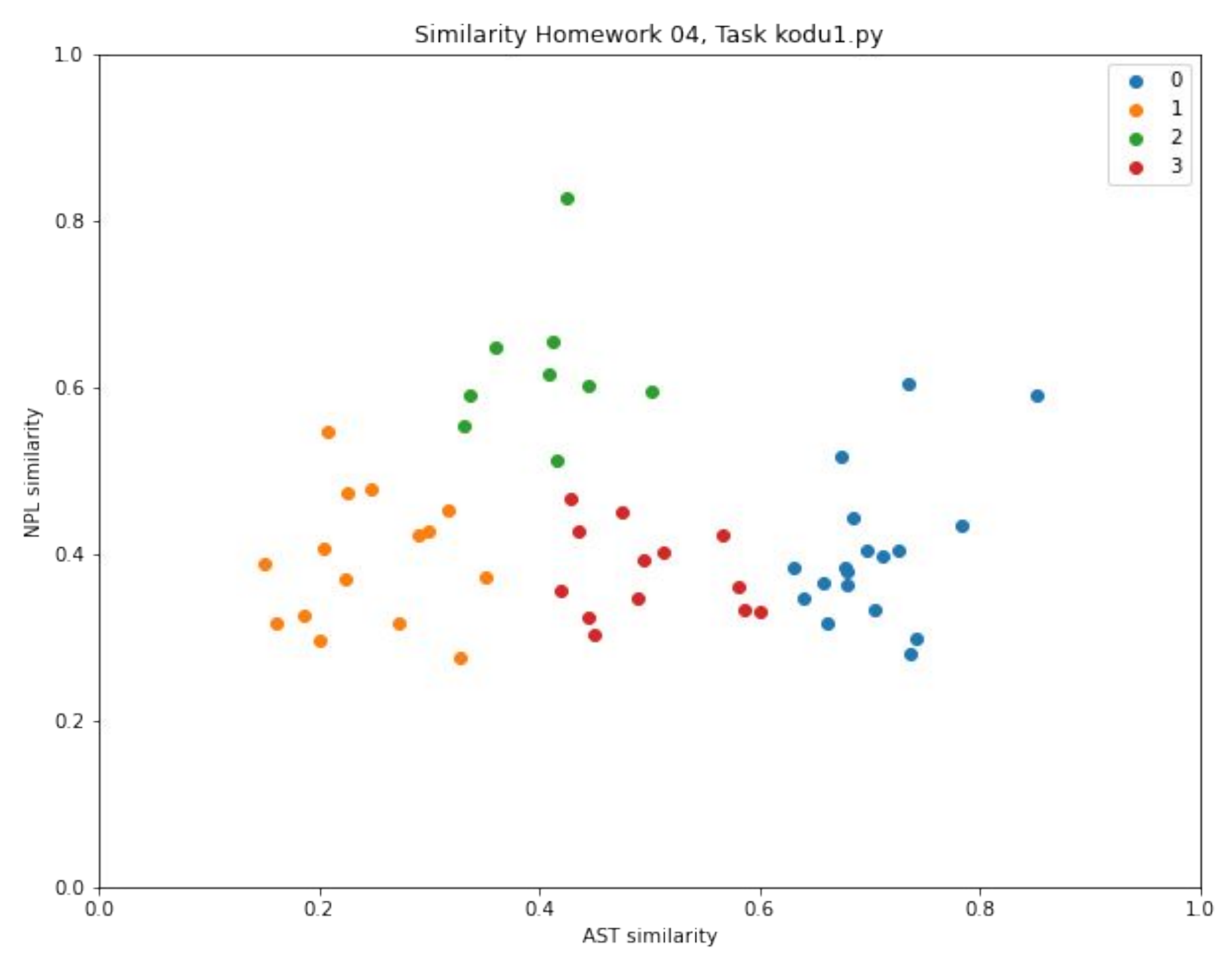
CLUSTERING

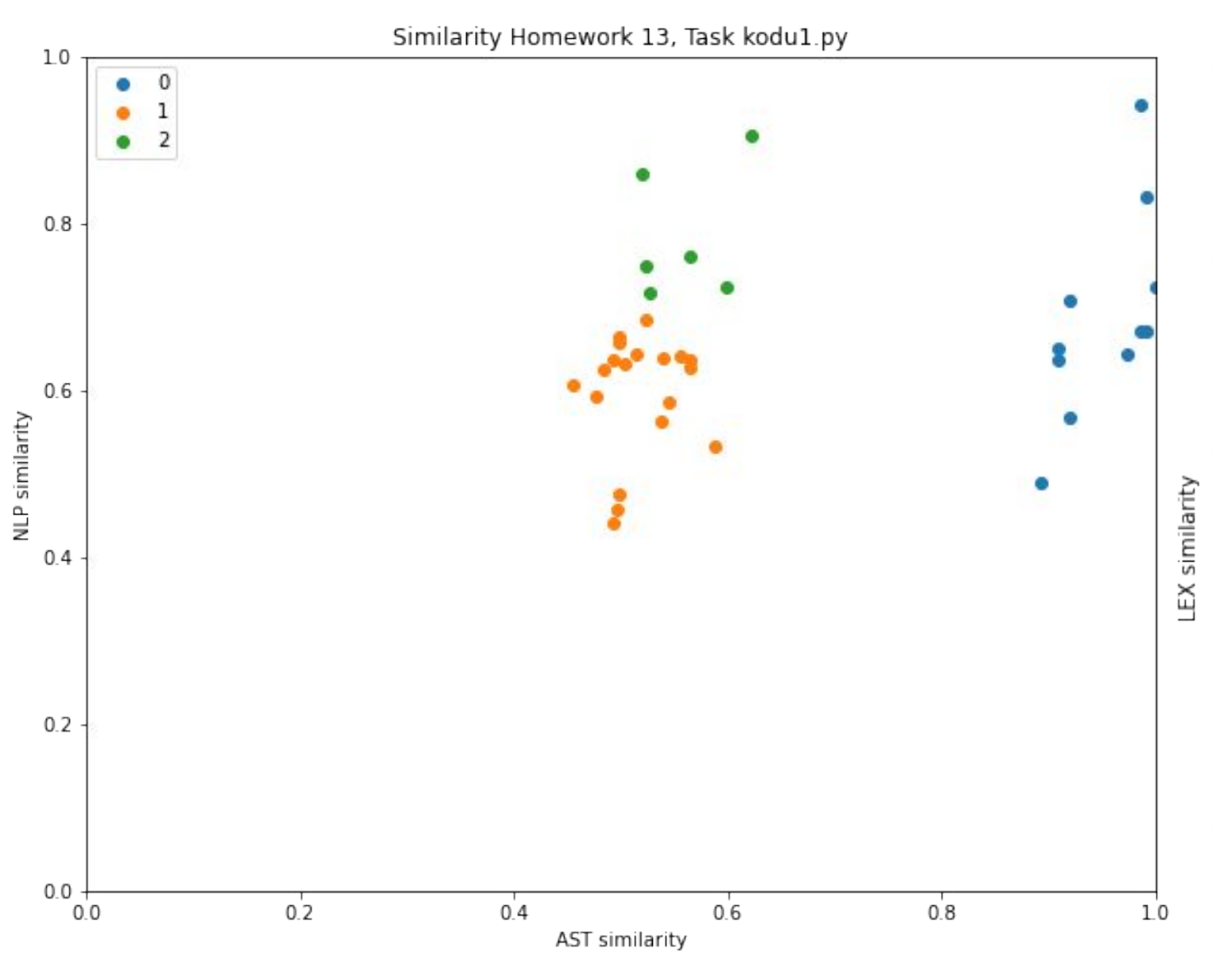
For clustering, we used K-means with Lloyd's algorithm and Euclidean distance to calculate the distance between two points. We clustered the points according to their lexical similarities using cosine similarity and abstract syntax tree similarities.

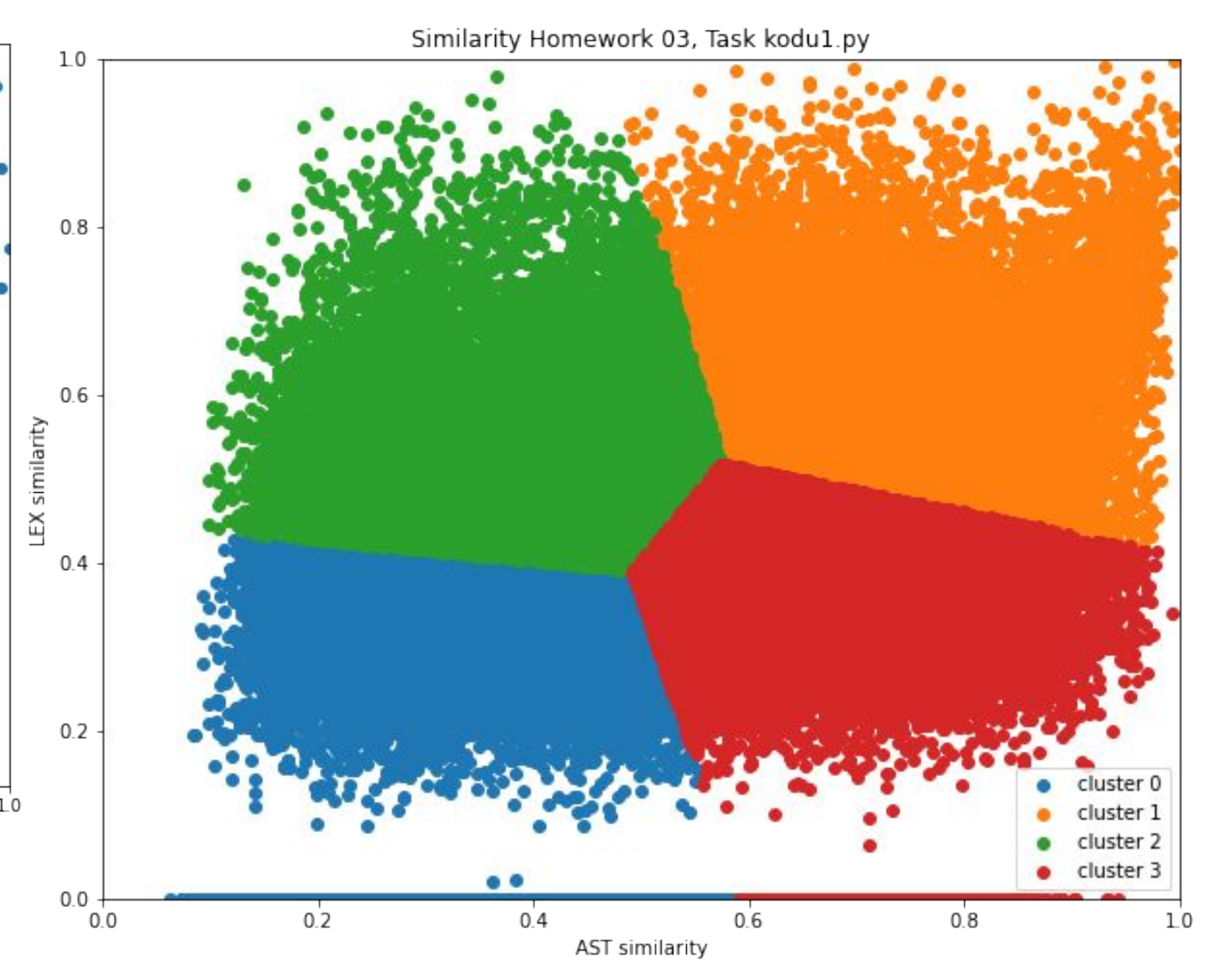
METHODS

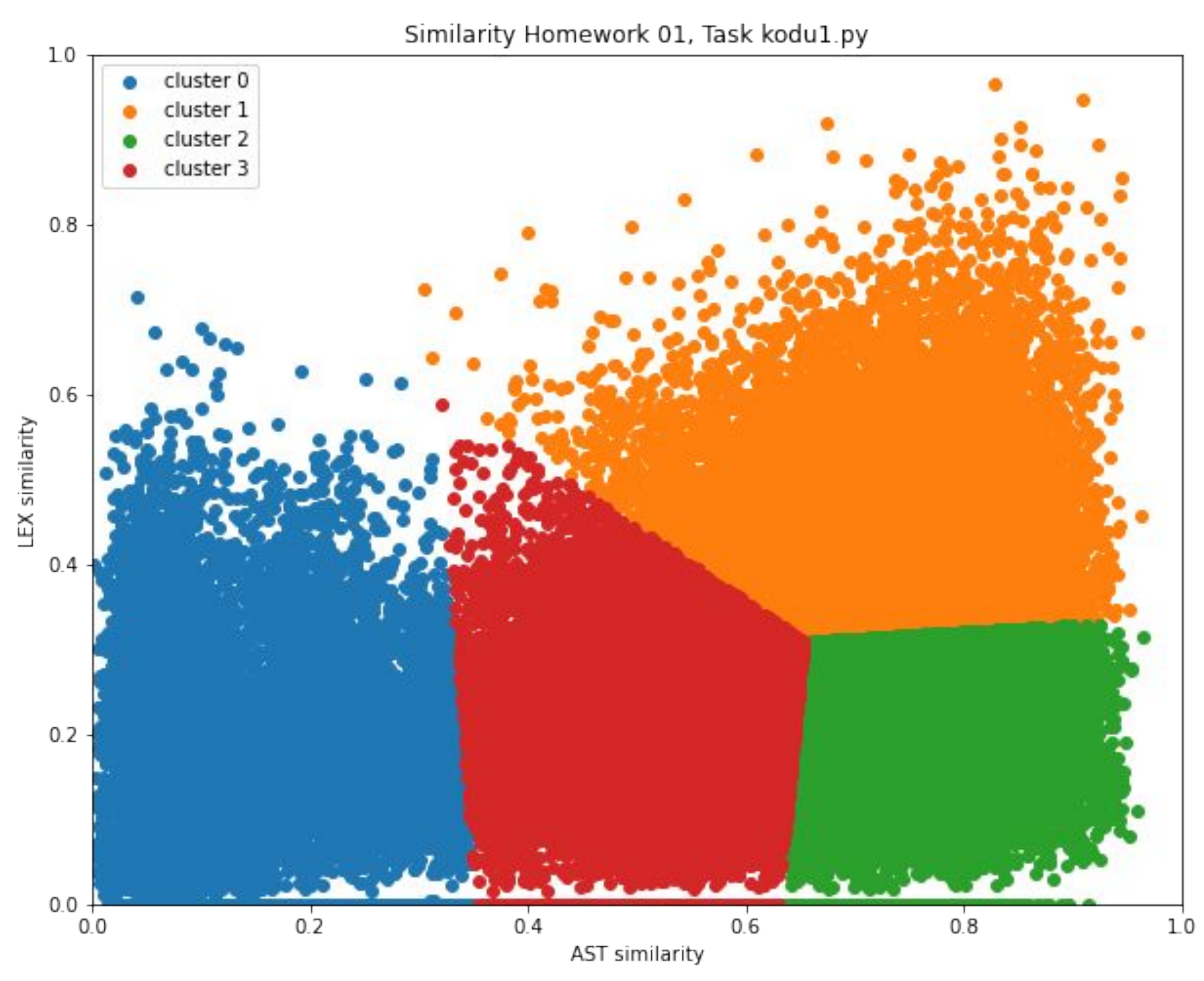
We used lexical analysis to detect similar wording, but this alone is not enough because you can hide collaboration by changing variable names, so we also used abstract syntax trees which convey the structure of the source code and increase our chances of finding similar programs

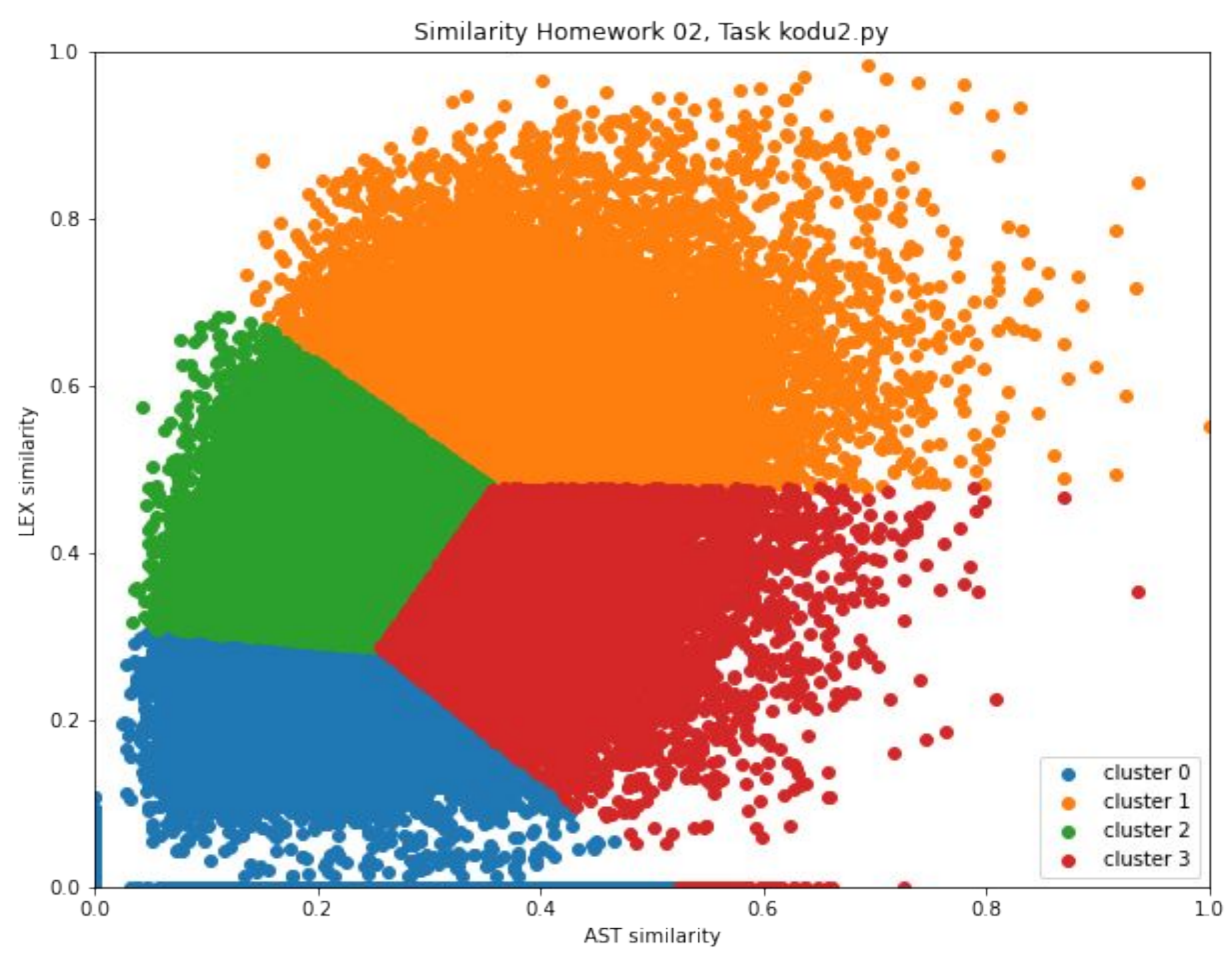


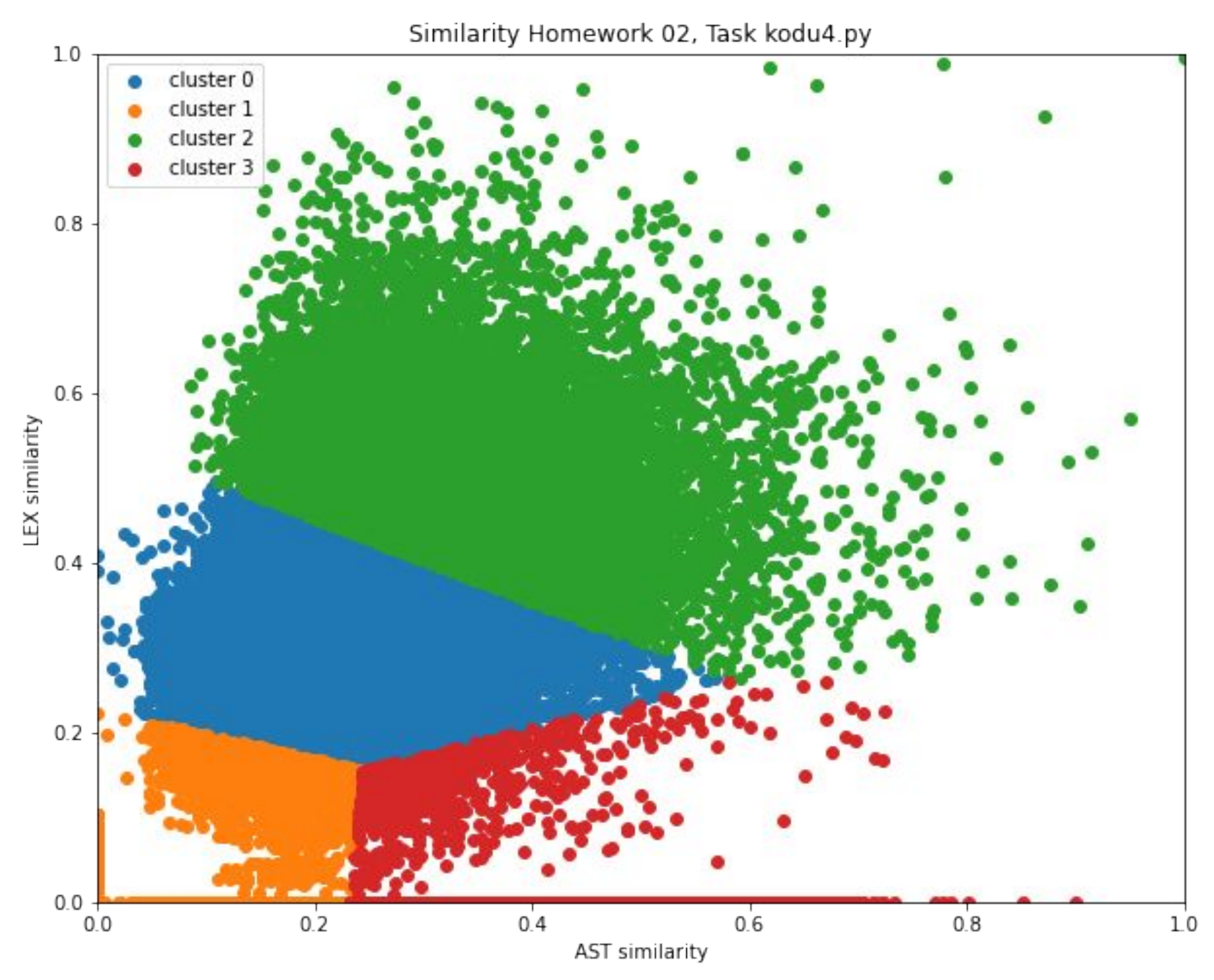






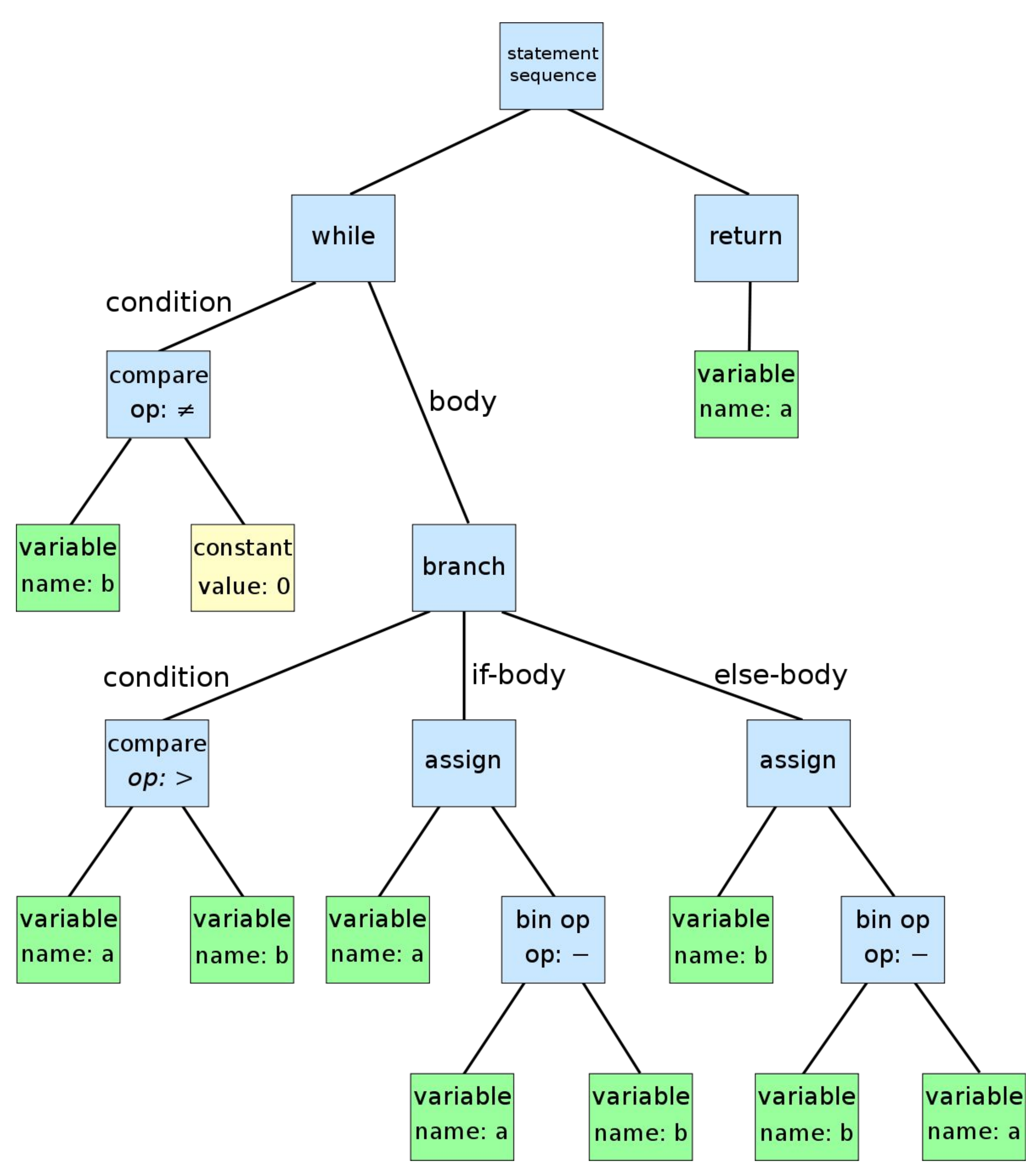






RESULTS

- The higher the AST similarity the more probable it is that there is collaboration or plagiarism involved
- Lexical similarity alone is not enough
- K-Means might not be the best clustering method for this sort of task
- Detecting collaboration and plagiarism in source codes can be a challenging task because programs can have the same lexical similarity but do completely different things and abstract syntax trees work if there are multiple approaches because as in Homework 13 kodu1.py we can see that the solutions have a fairly high similarities in both aspects.



https://upload.wikimedia.org/wikipedia/commons/c/c7/Abstract_syntax_tree_for_Euclidean_algorithm.svg

GitHub: https://github.com/Aksel2/IDSPython