Clustering Python programs

Team members: Mykyta Voievudskyi, Artur Aleksander Kanošin, Nikita Kislõi
Repository link: https://github.com/ArR4e/DSProject

Business understanding
Identifying your business goals
Background
Our client, Reimo Palm, PhD in Computer Science and a lecturer at University of Tartu, seeks to enhance the quality of the teaching at the Programming I course.

As every other course in university, the Programming I course has a list of compulsory homeworks. These homeworks are meant to master skills acquired at lectures and practical sessions; therefore, these homeworks are meant to be solved individually and just copying someone else's homework is strictly forbidden and considered an academic fraud. The course is of paramount importance when it comes to understanding basic programming concepts which are widely used not only in Computer Science but in other different fields as well. Thus, it is especially important to indicate academic fraud, particularly in this course.

The target audience of this course is beginners who had little to no previous experience with programming and Python in particular. For that reason, it is important to make the course less stressful for the students. It can be achieved by optimising tests: making them more flexible to accept different solutions that are correct in the context of the given task.

Business goals
To achieve what has been discussed so far, the following business goals were formulated:
Goal 1: For each task, it is required to find clusters of similar programs. This is meant to help to detect collaboration and discover different approaches to the task.
Goal 2: For each task, it is needed to find the outliers. This would improve the automatic testers used in the course for assessing the student programs.

Business success criteria

For our project to be successful the following criteria should be met: clusters of similar programs along with outliers are detected. Ideally, the model should detect as much collaboration and outliers as possible, but detecting 4 out of 5 cases of plagiarism and improving 7 tests is a lower bound for success.

Assessing your situation

Inventory of resources:

- Reimo Palm is a goto person, should any business related question arise.
- The dataset required for the modelling is provided by the client.
- Visual Studio Code is the IDE of choice for this project.
- Python 3.10, its built-ins as well as its libraries: NumPy, pandas, scikit-learn etc. will be used to develop the project.
- Jupyter is backend for writing jupyter notebooks. The source code will presented in jupyter notebook
- GitHub is where the source code will be hosted.

Requirements, assumptions, and constraints

The delivery of the project is scheduled for the 12th of December. Another requirement(proposed by authors) is that the model would minimise the cases when model reports a false positive plagiarism. The main assumption(used for optimisation) is that submissions from the same person are similar enough to consider only the last one.

Risks and contingencies

The main risk is the data leak. A set of successful solutions for homeworks is a sensitive data that should by no means be shared to anyone except for interested parties.
In addition, it is also undesirable to predict someone's homework as plagiarism when it is actually not and impose an immeasurable amount of stress on a student.

Terminology

AST – abstract syntax tree, represents syntax structure of source code, as seen by interpreter or compiler.
Token – basic component of source code.
Cluster – grouping of similar objects.
Accuracy – ratio of correct predictions to all predictions.

False positive rate – proportion of false positive predictions to all negative predictions.

Costs and benefits

Since it is an academic project, cost benefit analysis is not applicable here.

Defining your data-mining goals

Data-mining goals

The project has following deliverables: this very report, model(the priority is to minimise the number of false positives and maximise the true positive ratio.), poster and poster presentation.

The given dataset consists of the python scripts that could have potentially been altered just by swapping some rows, changing variable names, changing input-output texts, and adding or removing comments while maintaining the general logic and flow. For that reason, it is also necessary to preprocess and clean it and extract a more general structure from the script. That can be achieved by deleting comments, considering the last submissions and transforming programs into more generic form by creating AST, abstract syntax trees, or sequence of tokens the program, both approaches will be considered.

Upon processing, the dataset will be used to create clusters.

Data-mining success criteria

Success will be measured based on accuracy and false positive rate. The model is hoped to have at least 70% accuracy.

Data understanding

Gathering data

Outline data requirements and *Verify data availability* points are irrelevant since the data the model will be based on has already been provided.

Define selection criteria

The data is kept on cloud. The last submission is selected from multiple ones per person.

Describing data

Data is accessible via owncloud as a nested archive of submitted scripts. Data is fully compliant with what we need. Dataset itself consists of programs grouped by homework number and task number; there are 13 homeworks with the toal of 33 tasks and about 300 submissions per task. Inside each file, the students are coded as S001, S002, etc., consistently across the weeks. For example, student S001 in week 1 is the same student as student S001 in week 2.

Under each student, the folders contain all the submissions of that student. The name of the folder shows the date and time of the submission. For example, the folder 2021-09-02-07-30-37 contains the programs that were submitted on September 2, 2021, at 07:30:37. The folders ending with '.ceg' contain technical data added by the VPL tool (testing results, grades, etc.) which is irrelevant to current project.

The files in each submission have standardised names: kodu1.py, kodu2.py, etc., which mean Task 1 of the homework, Task 2 of the homework, etc., respectively. In homework 8, the filename is film.py instead of kodu8.py. If some student has no file that the other students have, then this student didn't submit the program for that task.

It should also be noted that although the last number of homework is 15. There are indeed 13 distinct homeworks, since there is no homework labelled as 6th or 12th.

Exploring data

Data exploration is irrelevant since each submission is unique except for fraud. Each submission has a metadata provided with it (grade, time etc.) yet this metadata is not related to the project as stated above. The number of lines of code could also have been measured, but it also has no direct influence on the result of data mining.

Verifying data quality

The quality of data is good enough in our case. It can be collected and saved for future use. Moreover, we have all submissions of different students in one directory which refers to concrete homework. Therefore, it makes the process of analysing, comparing and clustering concrete homework easier.

Planning your project

The workflow is as follows:

1. Delete all comments from the code because they are not necessary (this is too specific data which requires a number of arguable decisions to rely on in a case of plagiarism).
2. The next step is to process the script either by parsing it into AST or tokenising it, both methods will be considered and concrete one or both will be chosen later.
3. Then the data analysis will be performed by finding similarities in the data we extracted in point 2.
4. The next step is to analyse the results and find outliers.
5. Optionally, clusters, outliers and detected collaboration will be visualised if it will be applicable to the end result.
6. The codebase of the project will be reviewed and documented.

Methods

Cluster analysis, token sequence pattern-matching

Tools

The same tools will be used as described in *Inventory of resources*.