Team: Aksel Õim, Ken Böckler

Team nr: **C16**

Name: Clustering Python Programs

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

Business understanding

Background

Reimo Palm seeks to find collaboration between students and discover unique solutions in submitted homeworks to improve automatic testers. The detection of collaboration is important because it helps to identify plagiarism among students, and since plagiarism has a huge drawback for the educational process it is important to detect plagiarized work. The discovery of unique solutions helps to find different approaches to certain problems and through finding those unique solutions it is possible to improve the current automatic testers that are being used in the introductory Computer Programming course.

Business Goals

The goal of this project is to improve the automatic testers for Python homework tasks in the introductory Computer Programming course and detect collaboration among students.

Business success criteria

Successful results will mean that we have improved the automatic tests with our project and detected possible collaborations.

Inventory of resources

The main usable software for this project will be Python 3 and Jupyter Notebook. For data we will be using python programs that are divided between 13 homeworks and 33 tasks, the size for each program will be around 600 B and the total size for our dataset will be about 2340 KB. People working on this project will be 2 students, who can ask teaching assistants for guidance.

Requirements, assumptions, and constraints

The project must be finished by 12.12.2022 at 12:00. The requirements for the finished project are that the program must be able to find outliers and detect collaboration in Python programs.

Risks and contingencies

One of the biggest risks is losing the project to hardware failure, the solution to this is to have a repository that is up-to-date with the changes in the project.

Another risk is corrupting the data, the solution is to have a backup of the original data that was provided.

Internet outage in a home would be a problem because you could not keep the project up-to-date, the solution would be to go to Delta building until the issue is resolved.

Terminology

Clustering - Clustering algorithms find groups of items that are similar, in our case we are finding similar Python programs by clustering. Clustering divides a data set so that programs with similar content are in the same group, and groups are as different as possible from each other. This is an unsupervised learning method.

Data cleaning - The process of preparing raw data for analysis by removing bad data, organizing the raw data, and filling in the null values. Ultimately, cleaning data prepares the data for the process of data mining when the most valuable information can be pulled from the data set.

Data analysis - The process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

K-means clustering - is an unsupervised learning technique to classify unlabeled data by grouping them by features, rather than pre-defined categories. The variable K represents the number of groups or categories created. The goal is to split the data into K different clusters and report the location of the centre of mass for each cluster.

K-medoids algorithm - is a clustering approach related to k-means clustering for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data points in the cluster. These points are named cluster medoids.

Costs and benefits

This is not relevant to our project.

Data-mining goals

The data-mining goals for this project will be that we have made a poster presentation that showcases our findings, a written report that covers our work process and results in detail, and we have a processed dataset.

Data-mining success criteria

To support our business success we must have accurate clusters that effectively find Python programs that are outliers, and also the clusters must group programs that are similar to each other.

Data understanding

Gathering data

a. **Outline data requirements**

   Data must be in text type and qualitative (categorical). Our data is nominal. We can't compare our data to what is less or bigger but we can see if it's equal to. Maybe we need also homework description to do some categorization.

b. **Verify data availability**

   We already got our data into our filesystem. Reimo Palm sent us the data. The data is in usable form, structured and described.

c. **Define selection criteria**

   We have file-based data as a data source. The file extension is python, but it is essentially a text file. Let's take the exact same homework and compare the homework of different people. We find frequent patterns and get matching results that we are going to cluster. There are 13 files, each file contains all submissions of all students (about 300 students) for one week's homework. The number in the filename denotes the week of the homework; there were no homeworks in weeks 6, 12 and 16. The files were downloaded from the Moodle VPL tool.

Describing data

There are 13 files, each file contains all submissions of all students for one week's homework. The number in the filename denotes the week of the homework; there were no homeworks in weeks 6, 12 and 16. The files were downloaded from the Moodle VPL tool.

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.

Exploring data

The dataset and its structure are generally understandable. We select homework and people who have done this homework. Let's extract this dataset separately. We will make an automatic script for extraction. We put the dataset in a tabular form. Each row will represent one student. We have to think about classifying the columns. One way is to classify outputs not by code, but instead by code outputs at different stages of code execution. There is no false data in the data because we are comparing similarities.

Verifying data quality

We have studied the data and it is very good for our goals. We are missing some set data, but this does not prevent our work. In general, the data is very good and what can be improved becomes clear during the work.

Planning

0. Data gathering - Acquire the data needed for the project.

1. Data preparation - scanning the files to see if there is anything broken with the provided data, removing possible files that are corrupted or do not provide any value to the research, like empty files. To find similar programs from the submitted results we will be taking into account the last submission since the homework has multiple tasks and a task could have multiple submitted solutions and previously submitted solutions should be fairly similar to each other. To find outliers we will be using every submitted solution to increase the set of data and through that, we possible could increase the chance of finding an outlier. Time estimation: Aksel - 3h, Ken - 3h.
2. Decide which clustering method we will be using. We must select a clustering algorithm that is best suited for our dataset structure, this task includes research on the topic of clustering source code and other similar data. Time estimation: Aksel - 3h, Ken - 3h
3. Find the best way to read the data so that it could be used for clustering. We need to consider how we deal with comments variable names and having different input-output texts. Time estimation: Aksel - 2h, Ken - 2h
4. Prepare the data for clustering by having meaningful rows and columns. Time estimation: Aksel - 1.5h, Ken - 1.5h
5. Start writing the program which will be used for the clustering. This includes having graphs to visualise the findings. Time estimation: Aksel - 15h, Ken - 15h
6. Creating the poster for the poster session. Time estimation: Aksel - 4h, Ken - 4h

Right now we have decided to try K-means and K-medoids clustering algorithms. We will be using different Python libraries to ease the process of doing calculations by hand.