DOCUMENTATION

Marcin Kozub

March 2022

1 Steps in solving the task

Collecting files.

Firstly, I look for the CSV files that I can use to check how my program works during coding and test it later. I find them on kaggle.com.

- Employee Attrition: `www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data`
- Employee Future Prediction: `www.kaggle.com/datasets/tejashvi14/employee-future-prediction`

I think the data from the files is similar but I notice the problem in names of columns. I do not know if I should but I just rename the ones that are the same to have the exactly same names using PascalCase.

Idea of the algorithm.

After creating a new project and repository on GitHub, I finally start thinking about the idea of the algorithm in this program. I assume I cannot use libraries like Pandas, Numpy, csv etc... I do research and find some interesting facts about complexity. Let's assume that M, N are the number of lines in two CSV files. I know I can create $O(M \log M + N \log N + M + N) = O(M \log M + N \log N)$ time complexity algorithm if I load both files, sort them and join rows by iterating through them, but there can be a problem with loading too big files to the machine's memory. I don't want to load the whole files at once, I would like to read specific lines(rows) in both files instead. This approach costs me less in memory and solves loading too big files, but I write $O(N * M)$ time complexity algorithm.

Algorithm and tests.

For writing this task, I use pure Python. I check the input from the user. Test it and write the actual algorithm in my own created library for that task. Writing the code goes pretty smoothly, but it is interfered by classes that I need to attempt. I also created some smaller CSV files to test my program during writing it to be sure it works properly. I am not sure how the documentation should look like, then I decide to write doc strings in Python and don't copy them right here. I hope it's not wrong. The added tests help me to improve my

code and notice the mistakes I made. At the end, I don't know how to make my own command, so in order to execute my program you need to write:

python3 join.py file_path file_path column_name join_type