Lab 01

Data Preparation with Python

Task

Your task is to transform an existing data set into another format for further usage. Therefore you are going to write a simple Python program, which transforms the Netflix Prize data set from its proprietary format into a standard CSV format. The program may be as simple or as complex as you wish - but make sure you do not artificially increase the processing time. Try to make your program as performant as possible, as performance is one key element in the field of Big Data.

Steps

1. Install the Python (version >= 3) on your computer
2. Download the Netflix Prize dataset from Kaggle: https://www.kaggle.com/netflix-inc/netflix-prize-data
3. Extract the data and inspect the contents. You should find 4 files named "combined_data_x.txt" which contain the data we are going to work with
4. Write a Python application that is able to read in the data files and transform the proprietary format into the CSV format. Store your result as CSV file
5. Create an archive file (zip, rar, etc.) including your Python application and upload your solution to moodle (please do only include the program, not the transformed data)

Data Set Structure

The data includes the following fields:

- Movie ID (int)
- Customer ID (int)
- Rating (int)
- Date of Rating (date)

Unlike traditional CSV files, the movie id is not stored on every single line but separately.

```
1:
1488844,3,2005-09-06
822109,5,2005-05-13
885013,4,2005-10-19
30878,4,2005-12-26
823519,3,2004-05-03
```

The first line describes the movie id for all subsequent lines until a new movie id is defined. Movie id's are always denoted as integers followed by a colon. The other lines describe the rating per user: user id, rating, date of rating.

Output

- A python program to transform the Netflix Prize data set into a CSV file
- A single CSV file containing all data of the Netflix Prize data set in an appropriate format
- The input directory of the source files shall be parameterized as command line argument
- The output filename shall be parameterized as command line argument
- The program should be executed as
  - **python merge.py *<input_directory> <output_file>***

Outcome

- Basic understanding of the programming language Python
- Basic understanding of data transformation