Lab 02

Data Preparation with Python - Consistency Check

Background and Task

A very common task is to transform data to make it usable for other use cases. Do you remember the exercise from the previous Lab? You had to do exactly that. In business critical applications and environments it is immensely important that you can trust your data. In this Lab you will build a tool to check if what you have done in Lab 01 does produce the expected output.

Steps

1. Think about methods that can be leveraged to prove the consistency and completeness of the CSV output file of Lab 01. Find **at least two** independent checks.
2. Write a tool, e.g., in the form of a Python script, that reads a CSV file.
3. Implement the checks to make sure the data is processed as expected.
4. The tool should write into a file, “Data is beautiful” if all checks pass. If one or more checks fail, print “Data inconsistency found” and the reason for why the check(s) failed.
5. Run the script on your own Lab 01 output data.
6. Run the Lab 01 script of 2 other teams, run your Lab 02 script on the respective output data and let them know the result.
7. Upload the output (described below) to Moodle as a zip file.

Output

- A Python application to execute your consistency checks.
- A file with the result of your checks of your own output data named **results.txt**
- Each a file with the results of your checks of the other team’s data named **TEAMNAME_results.txt**
- The program should be executed as
    - **python check.py <input_file> <output_file>**

Outcome

- Deepen your Python programming skills
- Understand the importance of challenging the data you get from somebody else.
- Understand the importance of delivering high quality data.
- Understand that Lab02 is much more important than Lab01 in the business world.