Universität Göttingen
Department of Computer Science

Julian Kunkel

Exercise Week 7
HPDA / 2022
240 Minutes Total
Discussion in Week 8: 2023-12-11

---

1. The tasks described in this worksheet are part of the formative assessment. They serve the purpose to prepare you for the examination. We will discuss the solutions during the next **interactive session** after they are handed out – while they fit to the lecture of the week they are handed out, they might be discussed in two weeks time due to the bi-weekly exercise schedule.
2. Make sure to plan your time for the whole sheet carefully. The complete exercise should represent approximately three hours of independent study. The time limit indicates how much time you should spend on each task, and not how much time you may actually need; it is important that you engage with the material and not that you complete all tasks perfectly. Feel free to collaborate and team up.
3. The exercises are designed to challenge you and train you further as guided self-study. The time limit might be too ambitious for you; you may team up with colleagues. It is not an issue as long as you manage to at least partially resolve each task within the time budget. If you (and your team) are struggling, reach out for help in Teams! You may also share your thoughts via the Studip Forum.
4. We recommend that you create a (private) Git repository (see `https://gitlab.gwdg.de`) where you store your findings and outcomes while processing the exercises. This portfolio of work can be useful in the future.

Contents

Task 1: Spark Basics (120 min) **1**

Task 2: Interactive analysis of Bike Rentals (Spark with Jupyter Notebooks) (120 min) **2**

Task 1: Spark Basics (120 min)

As part of this exercise, we explore some Spark basics and perform some data inspection:

- First, download the data.
- Inspect the file briefly. The file may need some data cleaning to be ready to be loaded.
- Create an RDD by importing the CSV file.
- Use RDD operations to count the number of occurrences for each `crimedescr`.
- Save the `crimedescr` and frequency information into a CSV file.
- Compute the number for each `crimedescr` again. This time, use only accumulators.
- Now, use the methods for data frames for computing (particularly the SQL functionality).

Perform this task first using Spark's local mode in the virtual machine.

Portfolio (directory: `7/spark`)

`7/spark/crime.py` Your Python program performing all the tasks mentioned above

Hints

- You can replace the newline characters of Windows using `$ sed "s/\r/\n/g" <FILE>`.
- Further documentation about the operations from RDDs are provided at: http://spark.apache.org/docs/latest/.
- You may use *help()* in Python to start the Python help utility and retrieve the documentation for any Python module, class or function.

Task 2: Interactive analysis of Bike Rentals (Spark with Jupyter Notebooks) (120 min)

With Jupyter Notebooks, we can create interactive notebooks with embedded graphs that also use Spark!

- First, download the data.
- Run jupyter notebook
- Inspect the file briefly. The file may need some data cleaning to be ready to be loaded.
- Get started within the first cell as follows:

```
%load_ext autoreload
%autoreload 2

import os
import sys
# Load Spark python requirements into current scope
# TODO update directory:
spark_home = "/usr/local/lib/python3.8/dist-packages/pyspark/"

sys.path.insert(0, os.path.join(spark_home, 'python/lib/py4j-0.8.2.1-src.zip'))
sys.path.insert(0, os.path.join(spark_home, 'python'))
with open(os.path.join(spark_home, 'python/pyspark/shell.py')) as f:
    code = compile(f.read(), os.path.join(spark_home, 'python/pyspark/shell.py'), 'exec')
    exec(code)
import pandas as pd
print(sc) # Details about spark
```

- Plot a histogram for the duration of the bike rentals for durations smaller than 1 hour
- Use the RDD basic operations and alternatively SQL to extract the data

Portfolio (directory: `7/spark`)

`7/spark/bikes.ipynb` Your Notebook performing all the tasks mentioned above

Hints

- See the installation instructions for Spark in our GitHub repository
- You can register an RDD as a table using:
  `sqlCtx.registerDataFrameAsTable(table, "bikes")`
- You can plot data using:

```
import chart_studio.plotly as py
import plotly.graph_objects as go

# create a new plot
fig = go.Figure(
    layout=dict(
        title="Example Plotly plot",
        yaxis_type="log",
        xaxis_title='duration',
        yaxis_title='count',
    )
)

# Extracting data from the RDD df column ['col1']
data = go.Histogram(x=df.toPandas()['col1'])
# Plot the data
fig.add_trace(data)
fig.show()
```