

MapReduce

1. Bearbeite die folgenden Aufgaben mittels dem Programmiermodell MapReduce.
 - Achte darauf, dass Kopfzeilen korrekt behandelt werden.
2. Schreibe einen kurzen Report, der deine Lösungswege dokumentiert.
3. Reiche deine MapReduce-Programme als Python-Skripts und deinen Report als PDF gemeinsam in einer ZIP-Datei ein.
 - Achte darauf, deiner Abgabe keine Ausgaben (*.txt) direkt beizufügen.

1. Implementiere
für den Datensatz
mit folgender Ausgabe

und dem Ausgabeformat

[ProjectionDuplicateRemoval.py](#)
[geonames.csv](#)
Gib jeden Ortsnamen einmal aus, begleitet von der höchsten Bevölkerungszahl, die diesem Namen zugeordnet ist, falls er mehrmals in der Liste vorkommt.
Wien 1569316
2. Implementiere
für den Datensatz
mit folgender Ausgabe

und dem Ausgabeformat

[TopK.py](#)
[geonames.csv](#)
Gib für die $k(=20)$ bevölkerungsreichsten Orte in absteigender Reihenfolge jeweils die Bevölkerungszahl aus, begleitet von ihrem Namen.
6814400000 'Earth'
3. Implementiere
für den Datensatz
mit folgender Ausgabe

und dem Ausgabeformat

[MostCheckouts.py](#)
[checkouts-by-title.csv](#)
Gib jeden Autor einmal aus, begleitet vom Titel seines Buches, das am häufigsten ausgeliehen wurde, gemeinsam mit der Anzahl, wie oft es ausgeliehen wurde.
"Meyer, Stephenie, 1973-";"Twilight / Stephenie Meyer.",5750
4. Implementiere
für den Datensatz
mit folgender Ausgabe

und dem Ausgabeformat

[ECommerce1.py](#)
[2019-Oct.csv](#)
Gib für jeden Kunden die Marke aus, von der er die meisten Produkte gekauft hat.
"513600000" "samsung"
5. Implementiere
für die Datensätze
mit folgender Ausgabe

und dem Ausgabeformat

[ECommerce2.py](#)
[2019-Oct.csv](#) und [2019-Nov.csv](#)
Gib für jede Marke die Anzahl der Kunden aus, die sowohl im November als auch im Oktober etwas gekauft haben und die mehr Produkte dieser Marke als jede andere Marke gekauft haben.
"samsung" 13435