

Data Science 2Blatt 1, Abgabe am 23.10.2024 um 12:00

Geben Sie alle Dateien ab, die für das lokale Ausführen notwendig sind. Wenn Ihre Teststrategie (Aufgabe 2) kleine (!) Dateien erfordert, geben Sie diese mit ab. Wir werden jede Abgabe potentiell in den Übungen besprechen, daher sollten Sie keine personenbezogenen Daten in den Abgaben unterbringen.

Aufgabe 1. *Implementieren Sie Matrix-Matrix-Multiplikation für dünn besetzte (sparse) Matrizen im Coordinate-Format, also dass Einträge einer Matrix A als Tupel (i, j, A_{ij}) bzw. Dictionary-Einträge $(i, j) : A_{ij}$ gespeichert werden. Parallelisieren Sie die Rechnung mit MRJob.*

Wenn die Aufgabe zu groß erscheint: stellen Sie sicher, dass Sie es ohne Parallelisierung (wie auf dem vorigen Übungsblatt) implementieren können. Schauen Sie dann nochmal in das Beispiel zu word-count, wie MRJob benutzt wird. Ein weiteres Tutorial zu word-count mit MRJob ist hier zu finden:

<https://medium.com/databale/beginners-guide-for-mapreduce-with-mrjob-in-python-dbd2e7dd0f86>

Ansonsten ist auch ein Blick in die Dokumentation hilfreich:

<https://mrjob.readthedocs.io/en/latest/>

Sie sollten nicht versuchen, mit Hadoop, Spark oder einem Cloud Provider zu arbeiten - MRJob lässt sich vollständig lokal testen.

Ein anderer Weg, die Aufgabe schaffbarer zu machen: implementieren Sie als Zwischenschritt Matrix-Vektor-Multiplikation, wobei die Matrix dünn besetzt ist und der Vektor als 'voller' Spaltenvektor gespeichert wird. Das könnte etwas einfacher sein, auch wenn sich der so geschriebene Code nicht recyceln lässt für die eigentliche Aufgabe.

Aufgabe 2. *Entwerfen Sie eine (möglichst schlichte) Teststrategie, um Ihre Implementierung (und die der anderen Teilnehmer*innen) auf Korrektheit zu prüfen. Führen Sie die Strategie mit Ihrer selbst geschriebenen Methode durch.*

Aufgabe 3. *Reflektieren Sie, was Ihre Schwierigkeiten waren und wie Sie diese gelöst haben (oder was immer noch schwierig bleibt). Welche Fragen haben sich aufgetan und bleiben evtl. noch offen? Bei späteren Projekten sollen Sie einen Projektbericht schreiben, in dem Sie auch vergleichen, inwiefern der ursprüngliche Plan vom Endprodukt abweicht. Das ist hier wahrscheinlich nicht anwendbar, aber notieren Sie doch, was Sie überrascht hat und was Sie gelernt haben.*