

# Experimente gestalten fürs Maschinelle Lernen

## Übung 4: Multiprocessing und Projektplanung

Deadline: 07.12.2021, 9 Uhr

### Abgabe bei GitLab

Aufgabenteile gekennzeichnet mit **(A)** und **(B)**, müssen von unterschiedlichen Gruppenmitgliedern commitet werden (falls zutreffend).

## Multiprocessing

Ziele dieser Aufgabe sind:

- Das Programmieren zu üben. Am Code ist hier nichts kompliziert oder groß abstrahiert.
- Vertrautheit mit der `concurrent.futures` Multiprocessing Funktionalität zu gewinnen.
- Die Arbeitsweise mit Multiprocessing zu verstehen. Wird sehr nützlich, wenn Verarbeitung von Daten in ML Projekten viel zu lange dauern (mgl. Tage/Wochen) und die Zeit knapp wird. #deadline

**GitLab Portfolio** Arbeite weiterhin in Deinem GitLab Portfolio. Erstelle darin einen neuen Ordner `04_übung` und bearbeite den Multiprocessing Teil in diesem Ordner.

1. **(A)** Erstelle in `04_übung` einen neuen Ordner `multiprocessing`. Darin:
  - (a) Lade die Dateien `multi_at_home.py` und `multi_CAPS.py` von Moodle herunter und kopiere sie nach `04_übung`.
  - (b) Erstelle ein kurzes `README.md` für diesen Aufgabenteil. Erinnere Dich daran, was ein `README` in diesem Fall beinhalten muss (z.B. Requirements, Existenzzweck, Plan, TODOs). Das `README` darf gerne auch gegen Ende der Aufgabe die finale Form bekommen.
  - (c) Committe und pushe diese Änderungen nach GitLab. Erst danach, darfst Du mit dem nächsten Punkt weiter machen.

2. **(B)** Versuche `multi_at_home.py` auszuführen. Einige (mehr als nur ein) Fehler haben sich eingeschlichen. Arbeite Dich durch die Fehler durch und behebe diese.

Hinweis: Zeile 14 `for i in range(0, 1000):` sieht unnützlich aus, da führen wir die gewünschte Funktionalität 1000 mal aus. Das muss so und darf nicht verändert werden. Wir können uns damit vorstellen, dass der Text 1000 mal länger ist als unser kleines Beispiel.

3. **(A)** Ein böser Zauberspruch<sup>1</sup> hat den Skript aus der Vorlesung `multi.py` verändert (siehe `multi_CAPS.py`). `multi_at_home.py` ist unser Ansatz mit Python den Fluch rückgängig zu machen. Für das kurze Skript ist Multiprocessing etwas übertrieben, aber bei ganz großen Dateien (mehrere 100 MB) sehr sinnvoll. Probiere das gerne mit großen Textdateien aus, um Dich davon zu überzeugen.

---

<sup>1</sup>`tr a-z A-Z`

Erweitere kurz das Skript so, damit es das Original `multip.py` generiert und in den Ordner `04_übung` schreibt.

Hinweise:

- **Achtung:** Die Reihenfolge der Zeilen in `multip.py` muss deterministisch in der richtigen Reihenfolge geliefert werden (also bei jedem Durchlauf das gleiche, richtige Ergebnis und nicht durch Zufall).
- Ungefixte Bugs aus der vorherige Teilaufgabe müssen hierfür noch behoben werden. Einige Bugs verhindern nicht den Skriptdurchlauf. Damit ist konkret Zeile 16 gemeint. Wieso verändert man damit nicht die Liste? Behebe das und erkläre im `README` wieso Zeile 16 so nicht funktionieren kann.
- Die Aufgabe gilt nicht als gelöst wenn `n_worker = 1` gesetzt wird oder Multiprocessing ganz weggelassen wird.
- Die Aufgabe gilt nicht als gelöst wenn Zeile 14 `for i in range(0, 1000):` auskommentiert wird. Auf schnelleren Prozessoren reicht es diese Zeile auszukommentieren um *fast* deterministische Ergebnisse zu bekommen. Frage für Experten: wieso?

## Projektplanung

4. Erstelle einen neuen Ordner im GitLab Portfolio benannt `projekt` oder `project`. Erstelle ein `README` in diesem Ordner und schreibe darin einen Projektvorschlag indem Du folgende Punkte bearbeitest:
  1. Was ist das Problem was man lösen möchte?
  2. Aus welchen Daten kann/könnte man lernen? Hinweis: Bitte um Angabe des Links zu den Daten.
  3. Wie viele Trainingspunkte hat man zur Verfügung?
  4. Was sind die Features/Attribute? Hinweis: Bitte um eine Beispielhafte (tabellarische) Auflistung von etwa 6 solcher Datenpunkte mit Attributen.
  5. Mit welchen ML Algorithmen möchte man das lösen (komplett ausreichend: Decision Tree, Random Forest und Naive Bayes. Am liebsten alle 3 im Vergleich.)
  6. Wie kann man die Performanz des Modells messen?
  7. Was für eine Performanz auf das Problem wird erwartet?<sup>2</sup>

Falls Du keine Idee für ein Projekt hast, dann ist das Default Projekt etwas für Dich: Deine Task ist, aus diesen Daten<sup>3</sup> die Spalte "Rating"(TV Parental Guidelines) vorherzusagen. Mehr Details zu meinen Erwartungen dazu, kommen nach Eurem ersten Projektvorschlag Entwurf. Hinweis: Aus welcher Kombination von Spalten kann man das Rating am besten vorhersagen?

**Wichtiger Hinweis:** Normalerweise gilt, für alle anderen Blätter und Punkte, eine großzügige Nachreichungsfrist. **Punkt 4. aus diesem Blatt (Projektplanung) ist davon *ausgenommen*.** D.h., wenn die Deadline dieses Blattes für den Teil der Projektplanung ohne Abgabe des 4. Punkts überschritten wird, zählt es als fehlende Verpflichtung für ein Projekt, einen Vortrag und einem Referat.

---

<sup>2</sup>100% Performanz wird nicht in diesem Seminar erwartet!

<sup>3</sup><https://www.kaggle.com/shivamb/netflix-shows>