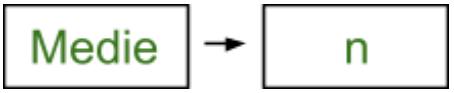
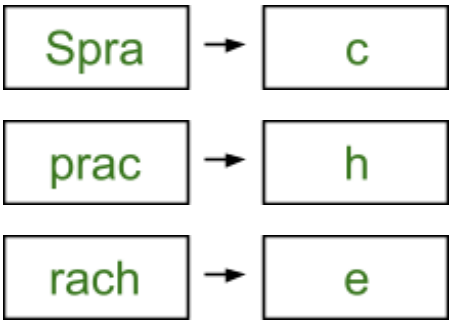
Fragebogen zur “littleLM”-Übung

1. Was bedeutet in der Übung "Zeichenfolge"?
   *Das Programm **schaut sich eine bestimmte Anzahl von Zeichen an**, um zu entscheiden, **welches Zeichen als nächstes kommen könnte**. Diese Anzahl von Zeichen bezeichnen wir in der Übung als Zeichenfolge. Eine 5er-Zeichenfolge besteht also aus 5 Zeichen.*

2. Stelle dir vor: Wir stellen eine 5er-Zeichenfolge ein. Wie würde unser Programm das Wort “Medien” zerlegen?



3. Wie würde das Wort “Sprache” bei einer 4er-Zeichenfolge zerlegt werden? (Hinweis: Das passiert in 3 Schritten.)



4. Welchen Einfluss haben Trainingsdaten-Größe und Zeichenfolge auf die Wortpaar-Übereinstimmung zwischen Trainingstext und generiertem Text?
   *Trainingsdaten-Größe: Je länger der Trainingstext ist, desto mehr Zusammenhänge kann das Programm zwischen den Zeichenfolgen und den nächsten Zeichen lernen. Wenn die Trainingsdaten dazu noch aus unterschiedlichen Textsorten/Themen bestehen, klingen die generierten Texte auch weniger nach den einzelnen Trainingstexten. Das bedeutet also, dass **die Wortpaar-Übereinstimmung 🔽 abnimmt, wenn die Trainingsdaten-Größe 🔼 zunimmt.***

Zeichenfolge: Je länger die Zeichenfolge - also je mehr Zeichen berücksichtigt werden, um das nächste Zeichen zu bestimmen -, desto sinnvoller klingen die Wörter und Sätze des generierten Textes. Allerdings klingen die Sätze dann auch ähnlicher wie die des Trainingstextes. Das bedeutet also, dass ***die Wortpaar-Übereinstimmung*** 🔼 **zunimmt, wenn auch die *Folge-Zahl*** 🔼 ***zunimmt.***

5. Was fällt dir auf, wenn du den Wert der Zeichenfolge
    a. niedrig einstellst (z.B. auf 2 Zeichen)?
    Die ***meisten Wörter*** *im generierten Text sind* ***"frei erfunden"*** *und haben keinen Zusammenhang zu benachbarten Wörtern.*
    b. hoch einstellst (z.B. auf 15 Zeichen)?
    (Fast) alle ***Wörter*** *im generierten Text* ***gibt es wirklich*** *(sofern sie im Trainingstext ebenfalls richtig sind). Wenn man allerdings den Trainingstext mit dem generierten Text vergleicht, fällt auf, dass* ***viele Sätze (fast) komplett kopiert*** *sind - es wurde also kaum ein neuer Text generiert.*

6. Was fällt dir am generierten Text auf, wenn du nur einzelne Trainingstexte auswählst? Mache mehrere Versuche mit unterschiedlichen Texten.
    Die Wörter und Sätze ähneln dem Trainingstext (stärker, je größer die Zeichenfolge). Wenn man einen anderen Trainingstext nutzt, hört sich also auch der generierte Text anders an.

7. Warum ist es besser, wenn im trainierten Modell für jede Zeichenfolge mehr als 1 nächstes Zeichen gefunden wird?
    Wenn jede Zeichenfolge nur eine Möglichkeit für das nächste Zeichen hätte, würde immer der gleiche Text generiert werden, da das Programm ja nicht zwischen mehreren nächsten Zeichen wählen könnte. Die generierten Texte wirken also mit mehr nächsten Zeichen pro Zeichenfolge kreativer.

8. Werden im Trainingstext mehr, weniger oder gleich viele Zeichenfolgen gefunden, wenn man die Zeichenanzahl erhöht? Warum?
   Eine längere Zeichenfolge kommt in einem Text meistens weniger oft vor als eine kurze Zeichenfolge. Zum Beispiel kommt die Zeichenfolge “lei” in den Wörtern “Blei”, “bleiben”, “klein” und noch vielen anderen vor. Die Zeichenfolge “nenunterga” vermutlich nur im Wort “Sonnenuntergang”.