

14. November 2022

Übungen 4

Bitte geben Sie nicht nur die Antwort an sich ab, sondern auch den R-Code, der Sie zur Lösung geleitet hat. Das gilt nur, falls die jeweilige Antwort R-Code zur Lösung benötigt.

01 *Alice in Wonderland*

- a. Installieren und laden Sie das Paket `{languageR}`.
- b. Schauen Sie sich die Elemente 16 bis 39 des Vektors `alice` an.
- c. Wie viele Worttoken (laufende Wörter) und Worttypen (unterschiedliche Wörter) sind in `alice` enthalten?
- d. Erstellen Sie eine Häufigkeitstabelle des Vektors `alice`, speichern Sie diese in einer Variable und geben Sie die 10 häufigsten Wörter aus.
- e. Erstellen Sie ein Balkendiagramm der Häufigkeiten der 10 häufigsten Wörter.
- f. Wie oft kommt das Wort "Rabbit" vor? Wie oft das Wort "rabbit"?
- g. Lassen Sie sich jedes Wort in `alice` ausgeben, das länger ist als 13 Buchstaben. Jedes Wort soll nur einmal in der Rückgabe auftauchen.
- h. Überführen sie den Vektor `alice` in Kleinschreibung, und speichern Sie das Ergebnis in einem neuen Vektor.
- i. Wie viele Worttypen sind in diesem neuen Vektor enthalten?
- j. Erstellen Sie eine neue Häufigkeitstabelle und lassen Sie sich wieder die Frequenz von "rabbit" ausgeben.
- k. Bonusaufgabe 1: Überprüfen Sie, ob die Summe von "Rabbit" und "rabbit" im originalen Vektor `alice` gleich der Häufigkeit von "rabbit" im neuen Vektor ist.
- l. Bonusaufgabe 2: Berechnen Sie normierte Häufigkeiten der Wörter, d.h. für jeden Worttypen, wie oft er in 1000 laufenden Wörtern vorkommt.

02 *Lexical decision*

- a. Lassen Sie sich Mittelwert und Median der Reaktionszeiten (Spalte `RT`) im Datensatz `lexdec` ausgeben.
- b. Nun lassen Sie sich erst den Mittelwert der Reaktionszeiten für Teilnehmer*innen ausgeben, deren Muttersprache (Spalte `NativeLanguage`) Englisch ist. Dann lassen Sie sich den Median der Reaktionszeiten für jene Teilnehmer*innen ausgeben, deren Muttersprache **nicht** Englisch ist. Sie können beide Mediane auch in einem Schritt anzeigen lassen.
- c. Wir möchten wissen, wie viele Wörter im Datensatz Tiere bzw. Pflanzen bezeichnen (Spalte `Class`) und wie sich diese Wörter über Simplexe (z.B. `dog`) bzw. Komplexe (z.B. `strawberry`) verteilen (Spalte `Complex`). Wir wollen

aber jedes Wort nur einmal zählen. Im Datensatz `lexdec` ist jedes einzelne Wort mehrfach vertreten, weil mehrere Teilnehmer*innen die gleichen Wörter verifizieren mussten. Beginnen Sie damit, einen Dataframe zu erstellen, der **nur** die Spalten `Word`, `Class` und `Complex` enthält.

- d. Unifizieren Sie diesen Datensatz. D.h. erzeugen Sie einen Dataframe, in dem jede Kombination aus `Word`, `Class` und `Complex` nur einmal vorkommt.
- e. Kreuztabellieren Sie die Spalten `Class` und `Complex` dieses unifizierten Datensatzes. Nun beantworten Sie die Frage: Wie viele Wörter entfallen jeweils auf die vier Kombinationen aus `animal/plant` und `simplex/complex`?