Tutorial 5: Kollokationen

Übung:

Für jede Frage geben Sie bitte ihren Code sowie die 10 höchstgerankten Kollokationen an.

1. Laden Sie die Datei “vb-nn-pairs.txt” von der Kurshomepage herunter, und lesen Sie sie in R ein. Sortieren Sie die Kollokationskandidaten nach gemeinsamer Frequenz, geben Sie die 10 häufigsten zusammen vorkommenden Wortpaare an.

2. Benutzen Sie den t-Test um die Kollokationskandidaten zu ranken. Als Gesamtanzahl der Bigramme rechnen Sie bitte mit 9266453. Geben Sie die 10 häufigsten und die 10 seltensten zusammen vorkommenden Wortpaare an.

3. Benutzen Sie den chi-square Test um die Kollokationskandidaten zu ranken. Geben Sie wieder die 10 häufigsten und die 10 seltensten zusammen vorkommenden Wortpaare an.

4. Benutzen Sie Pointwise Mutual Information um die Kollokationskandidaten zu ranken. Wiederum geben Sie bitte die 10 häufigsten und die 10 seltensten Wortpaare an.

TIPS:

- in R gibt es auch eine Funktion head().
- Sortieren in R: http://www.statmethods.net/management/sorting.html
- die Wurzelfunktion in R heisst sqrt().

VERA DEMBERG