R is a statistical and graphical programming language.

Run R by calling “R” in your terminal.

1. **Data types:** scalars, vectors (numerical, character, logical), matrices, data frames, lists. Here, we will only need Frames of numerical data. See http://www.statmethods.net/input/datatypes.html for examples of these data types.

2. **Reading in data** from a tab-delimited text file with column names:
   ```
   mydata <- read.table("mydata.txt", header=TRUE)
   ```
   set header=FALSE if you have a table without column names
   mydata is now of the type data frame.

3. **Viewing data**: a good way to view imported tables is using the summary function.
   ```
   summary(mydata).
   ```
   For numerical data, the summary function will show the min value, max value, mean, median, 1st and 3rd quantile of a data column, e.g. if our mydata table contains word frequencies, the summary under frequency will list the lowest and highest frequency, mean frequency, median frequency etc.

4. **Selecting** a particular part of the data in a frame:
   dim() tells us how many rows and columns the frame has.
   In order to select a particular enty, we can do frame[row-number,column-number], for example mydata[3,4] will return the value of the third row, fourth column in the data frame “mydata”. We can choose a complete column by underspecifying the row number: mydata[,4]. The same effect can be achieved via “mydata$ColumnName”, e.g. mydata$Length.

5. **Mathematical operations** can be called on single values such as log(1) = 0, or to whole vectors, e.g. log(mydata$Length) = (0 2.302585 2.995732 4.605170 7.170120)
   where mydata$Length = (1 10 20 100 1300).

6. **Plotting data:** http://www.statmethods.net/graphs/creating.html
   plot(vector-of-x-values, vector-of-y-values)
   A histogram:
   hist(vector-of-values)
   Saving graphs:
   call pdf(“mygraph.pdf”), then run the plot or hist command, then call dev.off()

For further reference see http://www.statmethods.net/index.html

Übungen:

1. Erstellen Sie basierend auf der in der letzten Übung erstellten Datei “german-word-freqs.txt” eine Datei, die den Frequenz-Rang, die Frequenz und das Wort token enthaelt, und speichern Sie unter “german-rank-freq.R.txt”, und benennen Sie die Zeilen als “Rank”, “Freq”, “Wort”. Die Zeilen in der Datei sollten durch tab getrennt sein, und es sollte keine Leerzeichen oder tabs vor der ersten Zeile geben.
   Beispiel: Wenn Ihre Wortliste 888 Wörter enthält, und das häufigste Wort “Katze” ist und 600 mal vorkommt, das zweithaeufigste Wort “Hund” ist und 342 mal vorkommt, und die beiden seltensten Wörter (hiervon wird es mehrere geben, Reihenfolge ist egal) “Dackel” und “Esel” sind und 1 mal vorkommen, muesste ihre Liste also wie folgt aussehen:

1	600	Katze
2	342	Hund
…	…	…
887	1	Dackel
888	1	Esel

   Tip: Sehen Sie sich mal die man-page des Kommandos *cat* an.

2. Laenge und Frequenz: Nehmen Sie die soeben erstellte Datei “german-rank-freq.R.txt” und fügen Sie eine weitere Spalte zur Länge eines Wortes hinzu. Diese Spalte soll den Titel “Laenge” tragen. Bitte speichern Sie die resultierende Datei unter “german-rank-freq-length.R.txt”.

1	600	Katze	5
2	342	Hund	4
…	…	…	…
887	1	Dackel	6
888	1	Esel	4

   Tip: perl hat eine Funktion length().

3. Starten Sie R und lesen Sie german-rank-freq-length.R.txt ein (falls Sie Aufgabe 2 nicht loesen konnten, lesen Sie einfach german-rank-freq.R.txt, die Spalte fuer die Laenge wird erst fuer Teilaufgabe 7 und 8 gebraucht). Überprüfen Sie, mit der Funktion “summary”, ob die Daten korrekt eingelesen wurden (i.e., achten Sie darauf ob die Zahlen unter Rank und Freq als Zahlen erkannt wurden und Ihnen die Summary Funktion die Verteilung mit min Wert, max Wert, Mittelwert etc. angibt). Falls nicht, überprüfen Sie, ob Sie die Titelzeile korrekt einlesen.

4. Plotten Sie Rank gegen Frequenz und speichern Sie den plot unter “rank-freq.pdf”

5. Plotten Sie log(Rank) gegen log(Frequenz). Speichern Sie den plotunter “logRank-logFreq.pdf”.

6. Erstellen Sie ein Histogram der Wortfrequenzen und speichern Sie unter “Hist-Freq.pdf”.

7. Erstellen Sie ein Histogram der Wortlaengen. Speichern Sie unter “Hist-Length.pdf”.

8. Plotten Sie log(Frequenz) gegen Laenge. Speichern Sie den plot unter “Freq-Length.pdf”.

9. Erstellen Sie ein Histogram des Verhältnisses von Wortlänge zu logarithmischer Wortfrequenz, speichern Sie unter “Hist-LengthFreq.pdf” und interpretieren Sie im Zusammenhang mit dem Plot aus Aufgabe 8.