# DBKDA-2016 Tutorial II:

# Hands-On Exercise 3

1. Load the webpage http://www.smiffy.de/dbkda-2016
2. Download the books „Moby Dick“, „Tom Sawyer“, „Ulysses“, and „Book from an unknown author“ to your local disk.
3. Evaluate the function ’readBookFromFile(...) (Appendix A) in your R-environmant.
4. Load the first 20 liones from the book „Moby Dick“ with the previously evaluarted function and examine the result.
5. Count the number of occurence for each word.
6. Draw a graph which shows the frequency of the most popular 20 words of the book Moby Dick (barplot).
7. Combine the bartplots from different books in one chart (StackedBarPlot)
8. Estimate from which author the book „Book from an unknown author“ was written.

- **Appendix A:**

```
# the function reads the given file (path) and returns
# a vector of words
# Parameters:
#    range: read lines from start:end (start:end)
#    stem: Stem the word, using the poerter stemmer
#          from the snowballC package
#
# example (read line 100 to 200 from the given file):
#    md<-readBookFromFile('c:/corpus/moby-dick.txt',
#                           range: 100:200)
#
readBookFromFile<-function(path,range=NULL,stem=FALSE) {
    lines<-readLines(path)
    if (is.vector(range))
        lines<-lines[range]
    lines<-tolower(lines)
    lines<-lines[lines!=""]
    words<-strsplit(lines, '\\W+')
    words<-unlist(words)
    words<-words[words!=""]
    if (stem)
     words <- unlist(wordStem(words, language="english")
    return(words)
}
# execute the following code to load the snowballC package:
if (! require('SnowballC'))
    install.packages('SnowballC')
library(SnowballC)
```

- **Some little Hints:**

1. To count the frequency of the words, try the function `table(...)`.
2. Use the  words as names for the columns (
   Example:

   ```
   names(vector)<- ...
   colnames(matrix)<- ...
   ```

3. To collect the words from different books try the union operator.
4. To sort by frequency use the `sort(..., decreasing=TRUE)` function call.
5. To compare the frequency of words from different books build a matrix, where the columns represent the different words and the rtows represent different books.
6. Ask us ..  ;-)