Statistical programming
Homework 3 – Friday, 28 October 2022
Due: Sunday, 11 November 2022

1. In the folder "bib" you have 50 files in RIS format, each containing information on one scientific work (e.g. journal article, conference proceedings paper …). From this you have to parse the following data (RIS tags are given in the brackets):
   a) Type of work (TY)
   b) All authors or the work (AU). Note that each work can have any number of authors. You should not consider other type of authors (those in other "tags", e.g. A1, A2 …)
   c) Year of publication (PY)
   d) Title (TI)
   e) All keywords (KW) of the work. There can be any number of keywords.
   f) Abstract (AB) if available (some works do not have it).

   The data must be saved in an R object that will allow you to easily process each field separately (e.g. all keywords), but it should also allow to extract all (or some) information for an individual work. This R object should be saved to a file (I suggest some of the R formats, preferably RDS (`saveRDS` , `readRDS`), but other formats that will allow you to read the data back are also allowed).

   You are not allowed to use any special packages that have support of RIS or similar format (but of course can use other string-processing packages).

   For this part, you have to submit the R code and the saved file.

2. Create a report (with R markdown) of the works, which will include:

   a. Overall number of works

   b. Plot representing the distribution of the type of works

   c. Some statistics and plots for:

      A. The length of the titles

      B. The number authors by work

      C. Publication year

      The statistics for all three variables must be presented in one table.

   d. The table or plot of most often appearing keywords. Keywords should be kept as they are, that is, they should not be split into separate words (if desired, some similar keywords can be merged, but this is not required).

   e. Some plot that will give some insight into the content based on Title + Abstracts. There are several ways you can do that, however in the simplest case (using the simplest case is perfectly OK), this can also be the table or graph of the 10 or 20 most common words with their frequencies. Wordclouds are also great. However, stopwords should be excluded. You can get all English stopwords with `stopwords::stopwords()`.

f. The rmarkdown file should use at least one rmarkdown parameter (e.g. the threshold for minimal frequency of words to display, the number of words to display, the name of the color palette).

g. All tables should be appropriately formatted (e.g. using kable). All tables and figures/plots must have captions.

h. For this part of the homework, you should submit the Rmd file, the knit version of the document (Word or pdf, but not html) and input data that are required (unless they were posted by me on in the e-classroom).