Kursus Bigdata F2017

Problem Week 8
Prepare R script files documenting the following:

Problem 8.1
In the following is focused on an initial experiment using the R packages "tm" for text mining and "wordcloud" for creating a selection of differently structured plots of frequently used sets of words, references [tm, 2016] and [wordcloud, 2016].

Verify the text mining experiment, implemented in the R code file "7_R_Text_Mining.R"

Notice that this experiment requires that the file doc_1.csv is in the RStudio working directory.

Study the following functions from the "tm" package:
?tm_map() and the selection of functions applied for cleaning the input text. The manual is accessed either from the RStudio console or alternatively via the reference [tm, 2016].

Modify moderately the R-code in "7_R_Text_Mining.R", with respect sizes of e.g. status plots, and verify that the modifications works correctly.

Problem 8.2
Now, use the above R-code for creating a wordcloud for each of the Coloplast A/S annual reports from 2011-12, 2012-13, 2013-14, and 2014-15. The annual reports are in pdf, so they are also distributed as *.csv in a format appropriate for this application.
Use 4 csv files for generating wordclouds for each of the annual reports. Finally verify manually if there might be differences between the annual reports, which can be identified in these 4 wordclouds, representing 4 consecutive years of annual reporting.
Notice: It might be useful to run a few experiments on how many words to include in the wordsclouds, e.g. 30, 50, 100 or maybe larger.

The resulting R code from Problem 8.1, 8.2 are inserted into a Problem8_xxx.R script file where xxx are characters chosen from the persons name. Each participant keeps the script for later submission.

Course material

[tm, 2016] https://cran.r-project.org/web/packages/tm/tm.pdf
[wordcloud, 2016] https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

2017.03.23/jaas