Introducing R Studio

September 1, 2017

In this course we will use a programming language called R for accomplishing machine learning. This might seem unfortunate since many of you probably have no or minimal knowledge of programming. However, you are not expected to have such knowledge or even to develop extensive programming skills during the course. We will only expect you to use already defined R *functions* and handle subsets of your data. Mikael and Johan will help you out if and when you get stuck.

Programming involves defining functions and procedures, which you will not do (but you may copy and paste functions from existing sources). For instance, calculating the mean value of the data in some table column C implies adding all values in the columns cells together and divide the sum by the number of table rows. This is not a necessary step to define, since we only need to issue the command `mean(C)`, thus making use of the already defined function `mean()` that takes a *parameter* that refers to your column data.[1] However, knowing how to grab the column of the complete data set is necessary.

The data you will be working with in this course and in the parallell Methods course mainly consists of plain texts or tabular data (numerical or categorical) in different forms. Most of the time, in the IR2/*Klassifikation and informationsutvinning* courses you will process *textual data* into different computable formats and in the Methods course you will mainly do calculations on subsets of *tabular data*. Moreover, you will visualize data in

[1] You will later learn that there are different datatypes that can be fed to functions, such as vectors, matrices, data frames etc

different ways in order to make them more susceptible to interpretation.

Our choice is R, since it is generally considered one of the best choices for data analysis, besides e.g. PYTHON. MICROSOFT EXCEL may be of help for some tasks but will quite soon come short for many reasons – however, you might want to use it (or OPENOFFICE's counterpart) for preparing data at some stage, which is possible. Moreover, *R* is our option even for the statistics part of the Methods course, in stead of MICROSOFT EXCEL or SPSS, which means that you do not need to learn different frameworks for different similar data analysis tasks.

When you have successfully installed the underlying R and R STUDIO you have access to an IDE (Integrated Development Environment) – i.e. R Studio – in which you are supposed to work with your data.

In Borås the IDE (R STUDIO) is already installed and you will find it in the menues. Just start it up by choosing that option. At home, you will find links to instructions for download and install in PingPong (or just turn directly to `https://www.r-project.org/`).

At the outset, we will lean on the work *Wickham, H. & Grolemund, G. (2017). R for Data Science. Sebastopol, CA: O'Reilly Media.* which is freely available on the web at `http://r4ds.had.co.nz/`, and refer to introductions and exercises in it. If you have the time before you come to Borås, read chapters One and Two. Continuing your work might be a good thing to do with this introductory book.

The time before lunch break in Borås we spend on the introduction is mainly aimed at getting you familiar with the environment, because many frameworks (which goes for MICROSOFT EXCEL as well) requires nearby assistance at the start of learning.[2] As soon as you have learned the basics, there are several ways to get help, but starting with teachers available in a computer lab is advantageous. *Always remember, you need to work tenaciously and indefatigably, because just like working with TEI, or any other XML application, problems may arise that needs time to solve – and this holds for experts as well.*

[2]If you are not present in Borås because you couldn't or wouldn't come you will of course have to manage on your own. For the Swedish group in the Methods course (who are not expected to participate on Thursday) the assignment tasks are more trivial and any problems can be handled by negotiating a time for an ADOBE CONNECT session.

Just as Wickham and Grolemund suggest, we will start exploring R by ignoring the problem of getting some data to work on and tidy it up. We will use data that is already present in R. By following the instructions given in **Chapter 3** you will get the opportunity to acquaint yourself with R just by repeating what is being done in the book (but also try the *Exercises* that are interspersed with the narrative). Just remember to have installed and loaded the necessary packages by the two commands

```
install.packages('tidyverse')
library(tidyverse)
```

after starting up the R STUDIO. The teachers will be of assistance during your work. Do as much as you have time to. Don't run to fast. It is better to think of what you are doing than just unconsciously repeating instructions. Ask questions, we will give you answers if we know them.

Exclusively for the IR2 and *Klassifikation och informationsutvinning* Courses

In this first introductory session you have been working with data that is already in R. However, as your project moves on you might create data on your own, either as tabular data or as texts.

In Wickham's and Grolemund's book this is covered in Chapter 10. Read and do the exercises in that chapter and try your skills on some data files provided in PingPong (or provided by some other means, if you are on the *Klassifikation och informationsutvinning* course).

Note that there are many ways in which you can import different data files.

Exclusively for the Methods Course

In the Methods course there are assignment tasks and exercises provided in PingPong. These are tailored towards other tasks than what the IR course implies, and not yet intended for those of you who are in the *Klassifikation och informationsutvinning* course – you will take this course next autumn.

Do these when you get home, after studying the statistics part of the Methods course and then observe that the instructions for some tasks, particularly data import, differ from those in the Wickham and Grolemund book.