«R Notebooks» og reproduserbarhet

Assignment 1 i kurset Data Science 2020

Skriv et kort notat — 5-7 sider (inklusive appendiks) i form av en «R Notebook» — der nødvendigheten av reproduserbarhet i forskning diskuteres. Diskuter også om bruk av «R Notebooks» er en mulig løsning på problemet med manglende reproduserbarhet.

Dokumentet må inneholde følgende bruk av R markdown:

1) Minst 4 overskrifter
2) Minst 1 ordnet liste på 2 nivå
3) Minst 1 eksempel på bruk av
    1) **halv-fet skrift(«bold»)**,
    2) *kursiv skrift («italic»)* og
    3) ***halv-fet kursiv skrift***
4) Minst 1 internt bilde skal være screenshot av git history som:
    1) Dokumenterer minst 10 «commits»
    2) Dokumenterer bruk av minst 3 «branches»
    3) Ekstra stjerne til dem som klarer å få til en «merge conflict» ;-)
    4) Bildet som dokumenterer git history skal være i et appendiks som kommer helt til slutt i dokumentet (etter referansene)
5) Kjør `sessionInfo()` i en code-chunk (husk å gi chunk-en navn). Hvordan kan denne funksjonen hjelpe oss med å gjøre et dokument reproduserbart?
6) Vi benytter apa for sitering og referanseliste (`apa-no-ampersand.csl` er tilgjengelig under *Filer* i Canvas.)
7) Bruk begge siteringsformene, dvs med og uten `[]`
    1) Husk at for å få siteringsinfo for R pakker kan dere bruke kommandoen `toBibtex(citation(<navn-R-pakke>))` , f.eks

```
toBibtex(citation("rmarkdown"))
```

```
## @Manual{,
##   title = {rmarkdown: Dynamic Documents for R},
##   author = {JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and K
##   year = {2020},
```

```
##   note = {R package version 2.3},
##   url = {https://github.com/rstudio/rmarkdown},
## }
##
## @Book{,
##   title = {R Markdown: The Definitive Guide},
##   author = {Yihui Xie and J.J. Allaire and Garrett Grolemund},
##   publisher = {Chapman and Hall/CRC},
##   address = {Boca Raton, Florida},
##   year = {2018},
##   note = {ISBN 9781138359338},
##   url = {https://bookdown.org/yihui/rmarkdown},
## }
```

Velg en «entry» — f.eks. fra `@Manual{`, t.o.m. `}` — vha. musen og kopier denne valgte teksten. Gå så inn i Zotero og velg `Importer fra utklippstavle` fra `Fil` menyen.

Forslag til litteratur

Se foredrag som ligger under *Kursets mediefiler* på Canvas. bib-filen som er brukt for referansene i foredrag er også lagt ut på Canvas (under *Filer*).

For generelle tanker rundt reproduserbarhet er Peng (2011) en god kilde. Videre gir McCullough et al. (2008) en god illustrasjon av problemets omfang innen fagområdet økonomi. McCullough et al. (2008) diskuterer også om tidsskriftenes arkiver av datasett og programkode er en tilfredsstillende løsning av problemet.

Basert på tanker fra Knuth (1992) introduserte Gentleman og Lang (2004) begrepet «compendium» som:

> both a container for the different elements that make up the document and its computations (i.e. text, code, data, ...), and as a means for distributing, managing and updating the collection.

Dokumentet nevnt ovenfor er det Gentleman og Lang (2004)[1] omtaler som «dynamic documents». Artikkelen drøfter også disse to begrepenes relevans for «reproducible research». Videre introduseres også «code chunks»

> sequences of commands in some programming language such as R or Perl. Code chunks are intended to be evaluated according to the language in which they are written. These perform the computations needed to produce the appropriate output within the paper, and also to produce intermediate results used across different code chunks.

[1] Robert Gentleman er sammen med Ross Ihaka regnet som «fedrene» til R

og «text chunks» som beskrives som:

> Text chunks describe the problem, the code, the results and often their interpretation. Text chunks are intended to be formatted for reading.

Disse tankene er også blitt brukt til å gjenskape deler av Golub et al. (1999) som et slikt «compendium» (Gentleman, 2005). Dette for å vise at idéen er gjennomførbar i prakis.

Når det gjelder «R notebooks», som kanskje kan betraktes som en implementering av et «compendium», er disse avhengige av de to pakkene `rmarkdown`, (Allaire et al., 2020), og `knitr`, (Xie, 2020). R markdown og tilhørende programvare er kanskje best beskrevet i Xie et al. (2018) og Riederer (udatert) .

Et forslag til disposisjon (dere trenger ikke dekke alt listet her)

- Innleding
  - Reproduserbarhet, R notebooks
- Litteraturgjennomgang
  - Replikerbarhet/reproduserbarhet
  - Problemets omfang
    - Vil dagens løsning med arkiv av data og event. programkode hos tidsskriftene kunne løse problemet?
  - Mulig løsning (teoretisk plan):
    - «Compendium», «Dynamic document», «code chunck» og «text chunk»
  - Mulig løsning:
    - R Notebooks
- Analyse
  - Løser R notebooks problemet med reproduserbarhet
    - helt eller bare delvis
  - Eksempler på «code chunks» («R Code Block») og «text chunck» i R notebook
  - Har forskerne incentiver til å være «reproduserbare», eller må de tvinges?
  - Er økt reproduserbarhet noe som vil tvinge seg frem eller er dagens økte interesse bare et blaff?
  - Kan reproduserbarhet ha relevans i sektorer utenfor akademia?
- Konklusjon
- Litteraturliste

R Notebook-en må kunne transformeres til .nb.html, .pdf fil. og MS Word format.

Litteraturliste

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., og Iannone, R. (2020). *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown

Gentleman, R. (2005). Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1034

Gentleman, R., og Lang, D. T. (2004). Statistical Analyses and Reproducible Research. *Bioconductor Project Working Papers*. https://biostats.bepress.com/bioconductor/paper2

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., og Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Knuth, D. E. (1992). *Literate Programming*. Cambridge University Press. http://books.google.com?id=vovpQgAACAAJ

McCullough, B. D., McGeary, K. A., og Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, *41*(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Riederer, E., Christophe Dervieux. (udatert). *R Markdown Cookbook*. Hentet 31. august 2020, fra https://bookdown.org/yihui/rmarkdown-cookbook/

Xie, Y. (2020). *Knitr: A General-Purpose Package for Dynamic Report Generation in r* [Manual]. https://yihui.org/knitr/

Xie, Y., Allaire, J. J., og Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown