R in Pairs

Wouter van Atteveldt & Kasper Welbers

Welcome to the “R in Pairs” workshop!

The goal of this workshop is to use R to make a word cloud of the pages in your canvas course. You should team up with someone (feel free to call them your “buddy” if that makes you feel comfortable) and work on the mini-project together.

We will be here to help, but always try to figure it out yourself first. You can use the 5 minute rule: if you feel like you’ve not made any progress in 5 minutes, ask us. Or just ask us whenever. We’re here for you!

For this workshop you can login to our hosted environment so you don’t need to install anything. However, you can install R and Rstudio for free on your own computer at any time if you want.

For more information, see our R material at https://github.com/ccs-amsterdam/r-course-material in general, and https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/R_basics_1_getting_started.md in particular.

Make sure to enjoy the chocolate, and take some with you to share and spread the good word! Remember: R is fun. And if you don’t think it’s fun, you’re not trying hard enough :)

(Why you should learn R? You can use it for statistics, but also for: visualization, data analysis/cleaning, network analysis, twitter scraping, text analysis, machine learning / deep learning, dealing with Canvas, and much more. And did we say it’s fun?)

1. Using R

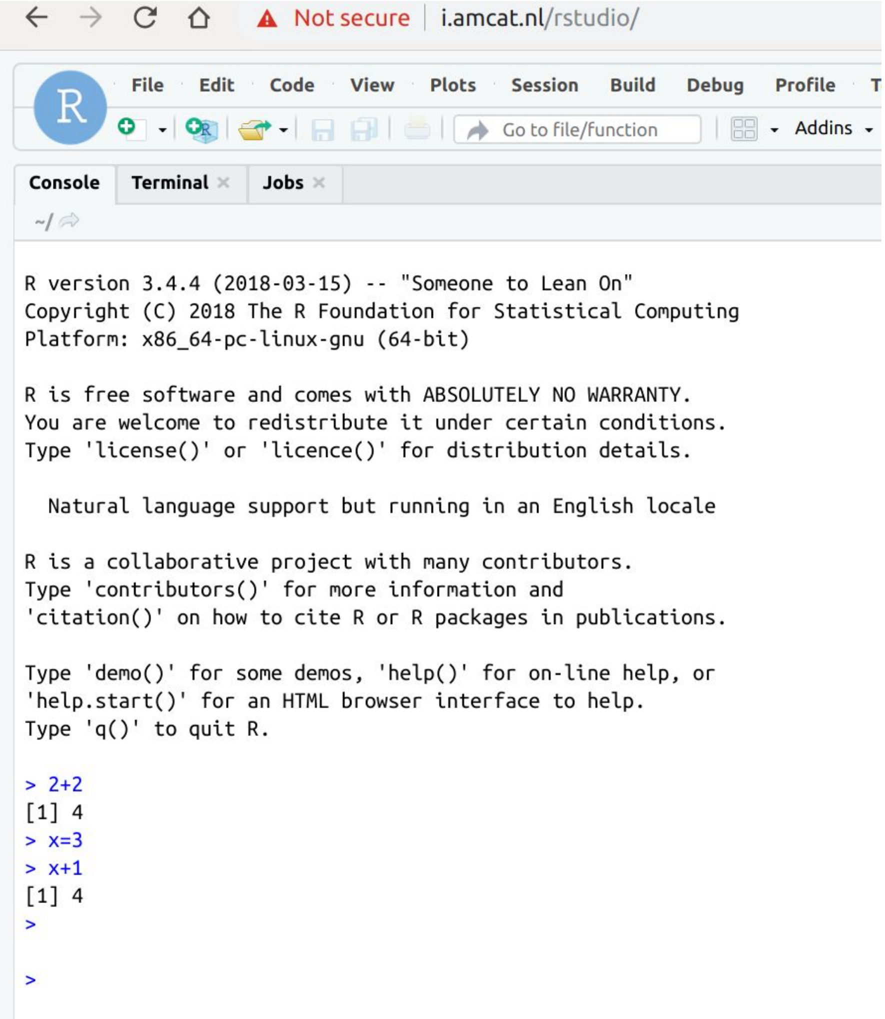
To get started, connect to our rstudio server at

http://i.amcat.nl/rstudio

Username: fsw01 - fsw05, password same as username. Every group should use their own login.

Now, you should be connected to R. You will see a ‘console’ on the left. You can execute a single command there, such as 2+2. You can also create variables using something like x=2.

Want to know more? See
https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/R-tidy-4-basics.md

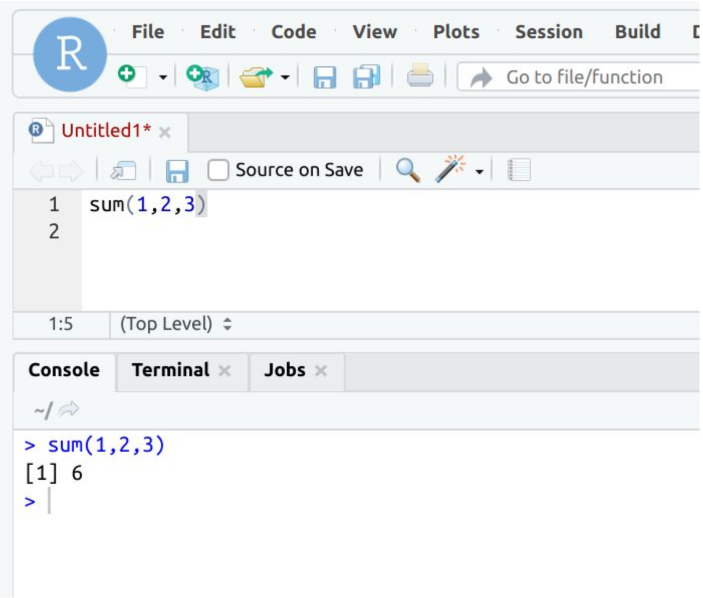


```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 2+2
[1] 4
> x=3
> x+1
[1] 4
>


>
```

2. Using Script files

Most of the time, we won't type commands directly in the console. Rather, create a script file using [File -> New -> R Script]. You can save and load this file later and this allows you to work on a data analysis script more easily than remembering to type all commands in the right order.

To execute a command, place your cursor in any line in the script and press control+enter to execute that line. You can also select part or all of the script and press control+enter or click on the "run" button. Try it by typing a command (such as 2+2 or sum(1, 2, 3)) and running it:



```
sum(1,2,3)
```

```
> sum(1,2,3)
[1] 6
>
```

3. Word clouds in R

For text processing (which includes word clouds), we will use the quanteda package developed at LSE. A package is a set of functions that was created by someone else and shared through the CRAN repository. Think of a package as an ‘app’ and the repository as an ‘app store’. If you need to install this (or most other) packages, for example because you are using R on your own computer, you can use the packages tab in the bottom right.

To activate a package, use the library command:

```
library(quanteda)
```

Now, you can use the functions from that package to make a silly word cloud:

```
text = "This is my text. It is like many other texts, but this one is mine"
d = dfm(text)
textplot_wordcloud(d)
```

Note that the quotation marks (") must be straight quotes, not the nice curly “quotes” that word often produces!

Note also that there are lots of options in the dfm command, such as for getting rid of HTML markup, strange characters, twitter symbols, numbers, etc.. You can also **dfm_trim** a dfm to remove all words that occur below a certain threshold. Textplot_wordcloud also has many options, from setting the word colours to changing how many words are displayed. To learn about these options, type the name of the function (e.g. **dfm_trim**) in the help pane in the bottom right. Also have a look at our material at https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/R_text_3_quanteda.md

4. Scraping Canvas

Of course, we'd like to get some actual texts to analyse. There are many different sources to use, such as political speeches (see 'getting started' link above), the Guardian or NYTimes newspapers (which both provide an API), our own AmCAT system (ask us :-)) or simply reading files from disk (see the readtext package, it is almost magical). But for this workshop, we will get data from canvas.

To do this, we use the rcanvas package (which is preinstalled on the server). You will need to get an API access token so the server know it's you:

1. Login to canvas
2. Select Account -> Settings
3. Click "new access token"
4. Type in a name (e.g. "R Workshop") and click Generate token
5. Copy the token (the long seemingly random text above "copy this token now")

In R, run the following commands to connect to canvas:

```
library(rcanvas)
set_canvas_token("YOUR_COPIED_TOKEN")
set_canvas_domain("https://canvas.vu.nl")
```

Now, you can download all your courses and their syllabus:

```
courses = get_course_list(include="syllabus_body")
```

Click the 'courses' variable in the top right 'environment' pane to have a look at what it downloaded.

5. Canvas word cloud

Now, we can combine sections [3] and [4] to create a word cloud of the downloaded syllabi:

```
d = dfm(courses$syllabus_body)
textplot_wordcloud(d, max_words=50)
```

That was easy, right? But not really pretty or informative yet. To improve this, let's first remove all HTML elements, and then use the dfm options to get rid of stopwords, numbers, and punctuation:

```
text = gsub("<.*?>", "", courses$syllabus_body)
d = dfm(text, remove_punct=T, remove=c(stopwords('nl'), stopwords('en')),
        remove_numbers=T)
textplot_wordcloud(d, max_words=50)
```

Better, right? You can also select a single course, e.g. Diversity 1:

```
select_course = courses$course_code == "S_D1"
d = dfm(text[select_course], remove_punct=T, remove=c(stopwords('nl'),
stopwords('en')), remove_numbers=T)
textplot_wordcloud(d, max_words=50)
title(courses$name[select_course])
```

Note that this will only work if you are part of D1, and it will also only work for courses that use the syllabus (which you really should!).

Finally, you can also download all announcements for a course, using the course ID from canvas or from the courses data we downloaded:

```
ann = get_announcements(course_id = 38555, start_date="2019-01-01",
                        end_date="2019-12-31")
ann_text = gsub("<.*?>", "", c(ann$title, ann$message))
d = dfm(ann_text, remove_punct=T, remove=c(stopwords('nl'), stopwords('en')),
        remove_numbers=T)
textplot_wordcloud(d, max_words=50)
```

You can also e.g. get the announcements, grades, users, group and much more, and you can also use this to post grades or change group settings (for example, to force specific group membership for each assignment). To see what else you can use rcanvas for, have a look at https://github.com/daranzolin/rcanvas. For more R material, check out https://github.com/ccs-amsterdam/r-course-material . Finally, be sure to check out https://www.r-graph-gallery.com/portfolio/ggplot2-package/ for some inspiration on visualizations. Have fun, enjoy the chocolate, and spread the word!