Polish Language(s) and Digital Humanities Using R

G. Moroz

2020

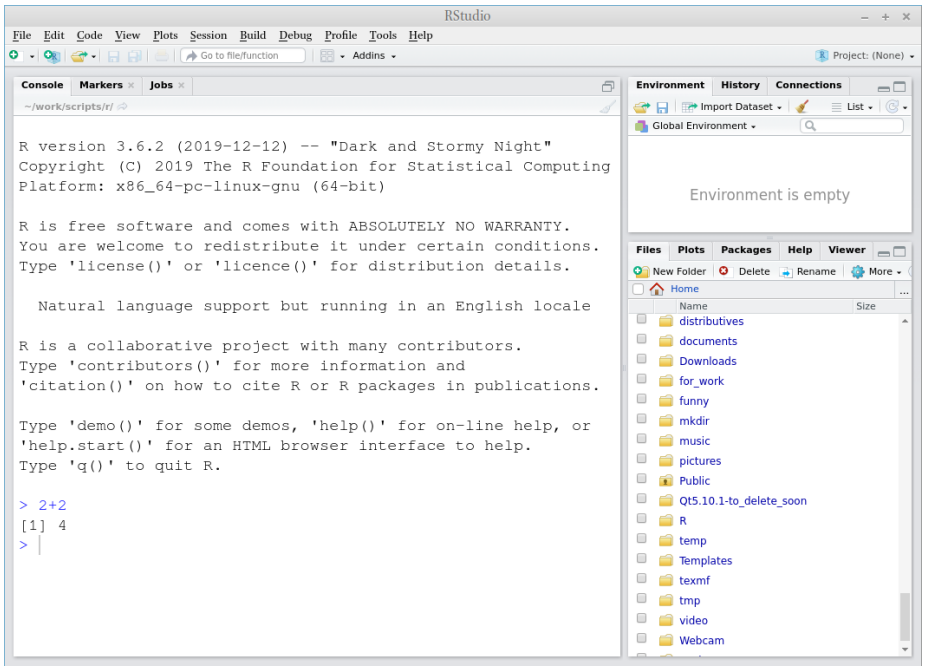
Contents

Chapter 1

Prerequisites

Before the classes I would like to ask you to follow the instructions mentioned below to prepare your device for the class work:

- install **R** from the following link: https://cloud.r-project.org/
- install **RStudio** from the following link: https://rstudio.com/products/rstudio/download/#download (FREE version, no need to pay!)
- after the installation run the RStudio program, type `2+2`, and press `Enter`.



If you see something like this, then you are well prepared for classes.

- Go to the https://rstudio.cloud/ website and sign up there. This is optional, but it will be a backup version, if something will not work on your computer.

Chapter 2

Introduction to R and RStudio

2.1 Introduction

2.1.1 Why data science?

Data science is a new field that actively developing lately. This field merges computer science, math, statistics, and it is hard to say how much science in data science. In many scientific fields a new data science paradigm arises and even forms a new sub-field:

- Bioinformatics
- Crime data analysis
- Digital humanities
- Data journalism
- Data driven medicine
- …

There are a lot of new books "Data Science for …":

- psychologists (Hansjörg, 2019)
- immunologists (Thomas and Pallett, 2019)
- business (Provost and Fawcett, 2013)
- public policy (Brooks and Cooper, 2013)
- fraud detection (Baesens et al., 2015)
- …

Data scientist need to be able:

- gather data
- transform data

- visualize data
- create a statistical model based on data
- share and represent the results of this work
- organize the whole workflow in the reproducible way

2.1.2 Why R?

R is a programing language with a big infrastructure of packages that helps to work in different fields of science and computer science.

There are several alternatives:

- Python (VanderPlas, 2016; Grus, 2019)
- bash (Janssens, 2014)
- java (Brzustowicz, 2017)
- ...

2.2 Introduction to RStudio

2.3 R as a calculator

2.4 Functions

2.5 Variables

2.6 Variable types

2.7 Vector

2.7.1 Vector coercion

2.7.2 Vector operations

2.7.3 Vector recycling

2.7.4 Indexing vectors

2.7.5 `NA` value

2.8 Dataframe (tibble)

2.8.1 Indexing dataframes

2.9 Packages

2.10 Data import

2.10.1 `.csv` files

2.10.2 `.xls` and `.xlsx` files

2.11 Rmarkdown

Chapter 3

Data manipulation: `dplyr`

Chapter 4

Data visualisation: `ggplot2`

Chapter 5

Strings manipulation: `stringr`

Chapter 6

Text manipulation: `gutenbergr`, `tidytext`, `udpipe`

Chapter 7

Stylometric analysis: `stylo`

Bibliography

Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Brooks, H. and Cooper, C. L. (2013). *Science for public policy*. Elsevier.

Brzustowicz, M. R. (2017). *Data Science with Java: Practical Methods for Scientists and Engineers*. O'Reilly Media, Inc.

Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media, Inc.

Hansjörg, N. (2019). *Data Science for Psychologists*. self published.

Janssens, J. (2014). *Data Science at the Command Line: Facing the Future with Time-tested Tools*. O'Reilly Media, Inc.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Thomas, N. and Pallett, L. (2019). *Data Science for Immunologists*. CreateSpace Independent Publishing Platform.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc.