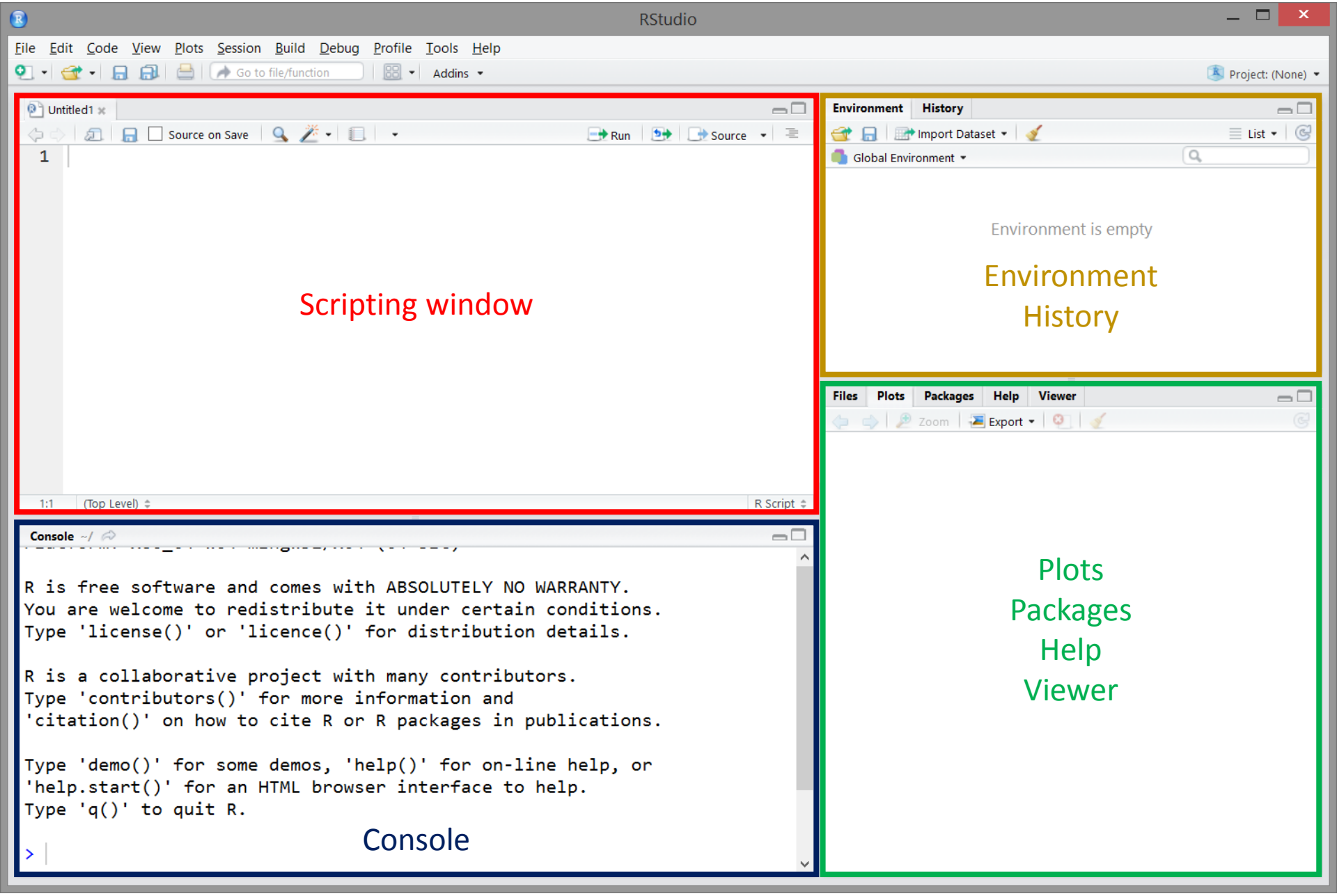
Text analysis 2

Lukáš Lehotský

R studio layout

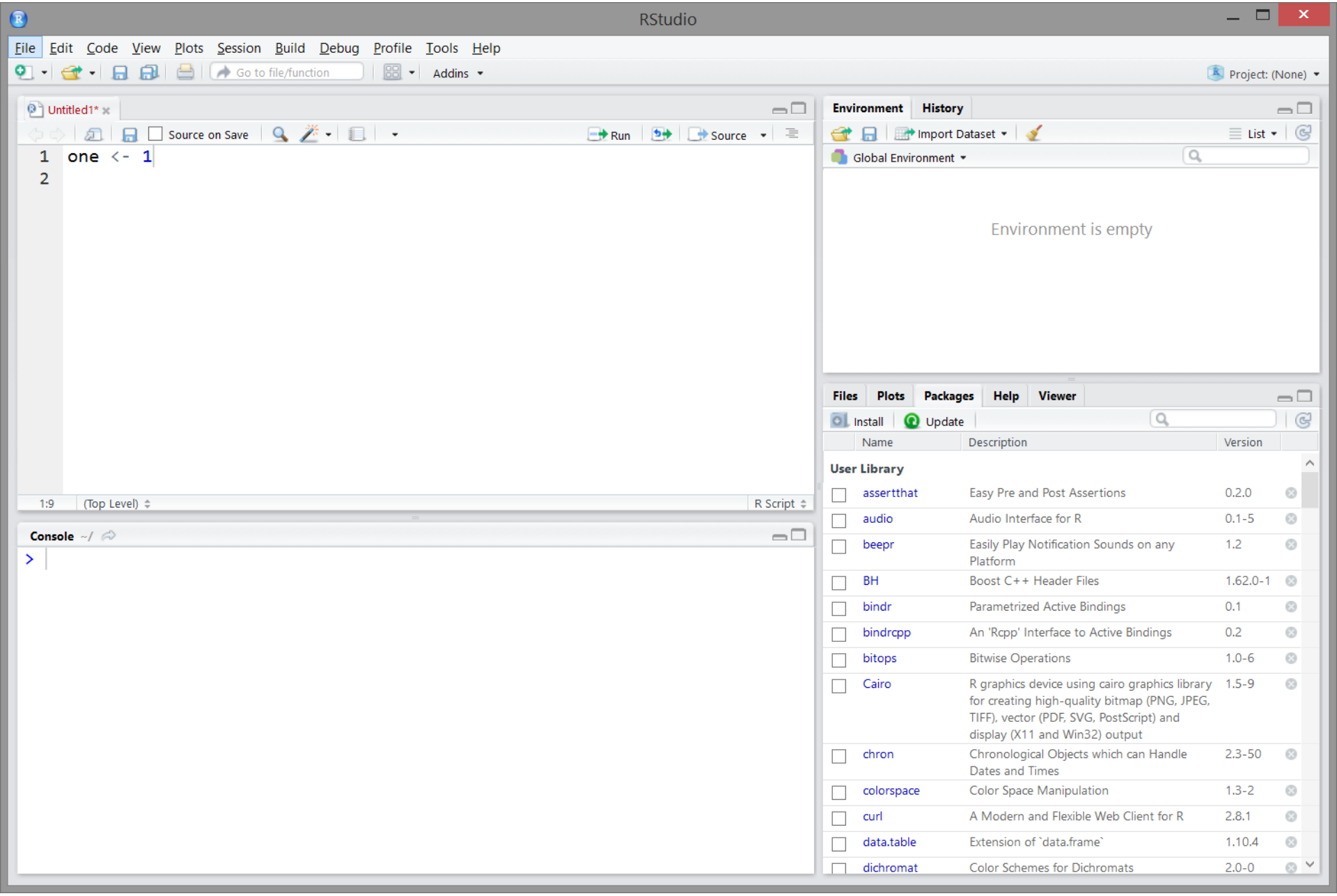
Scripting window	Environment (stored objects) History
Console window	Plots Packages Help Viewer

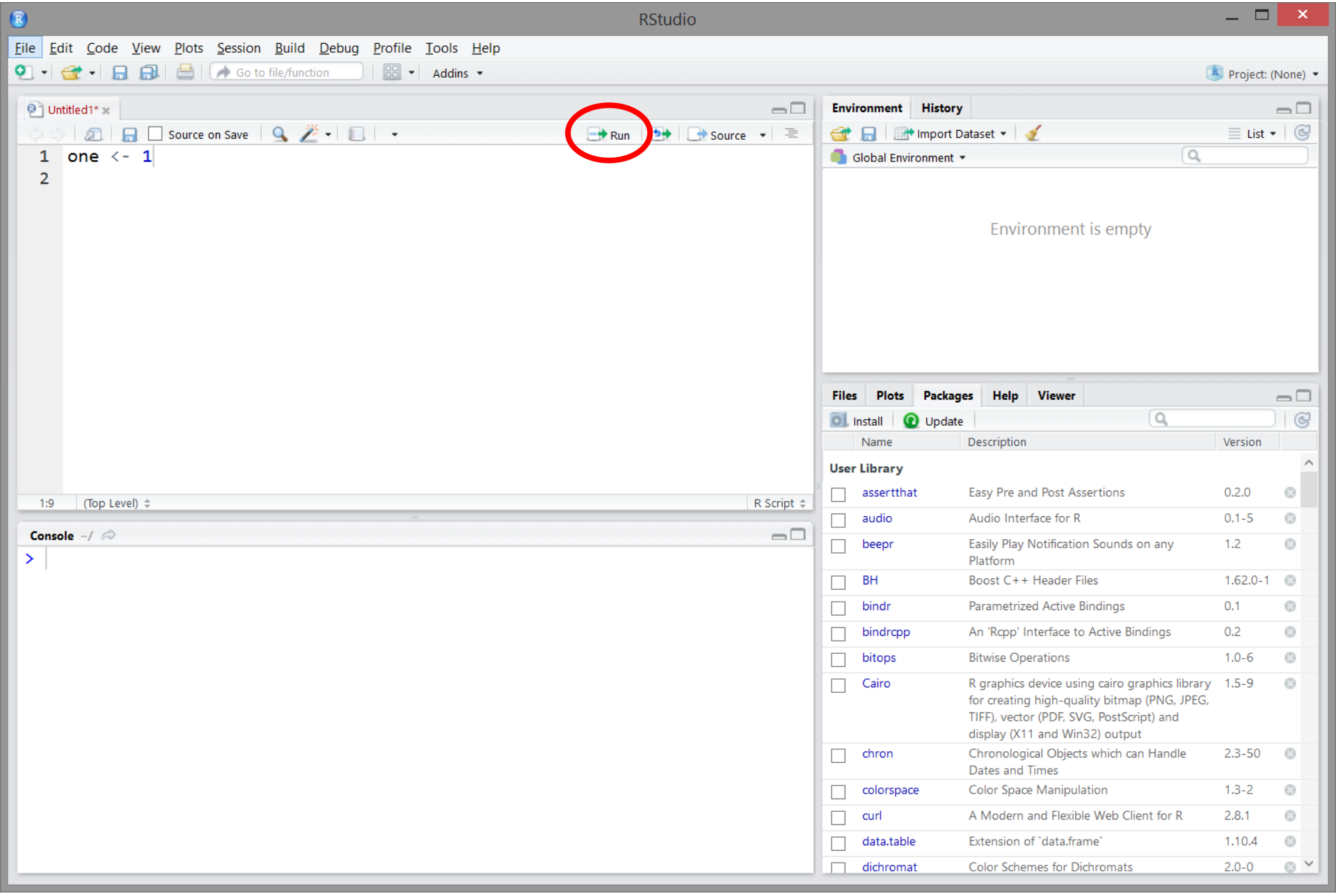

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
Scripting window
1:1
(Top Level)
R Script
Console ~/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
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.
>
Console
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Environment
History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Plots
Packages
Help
Viewer

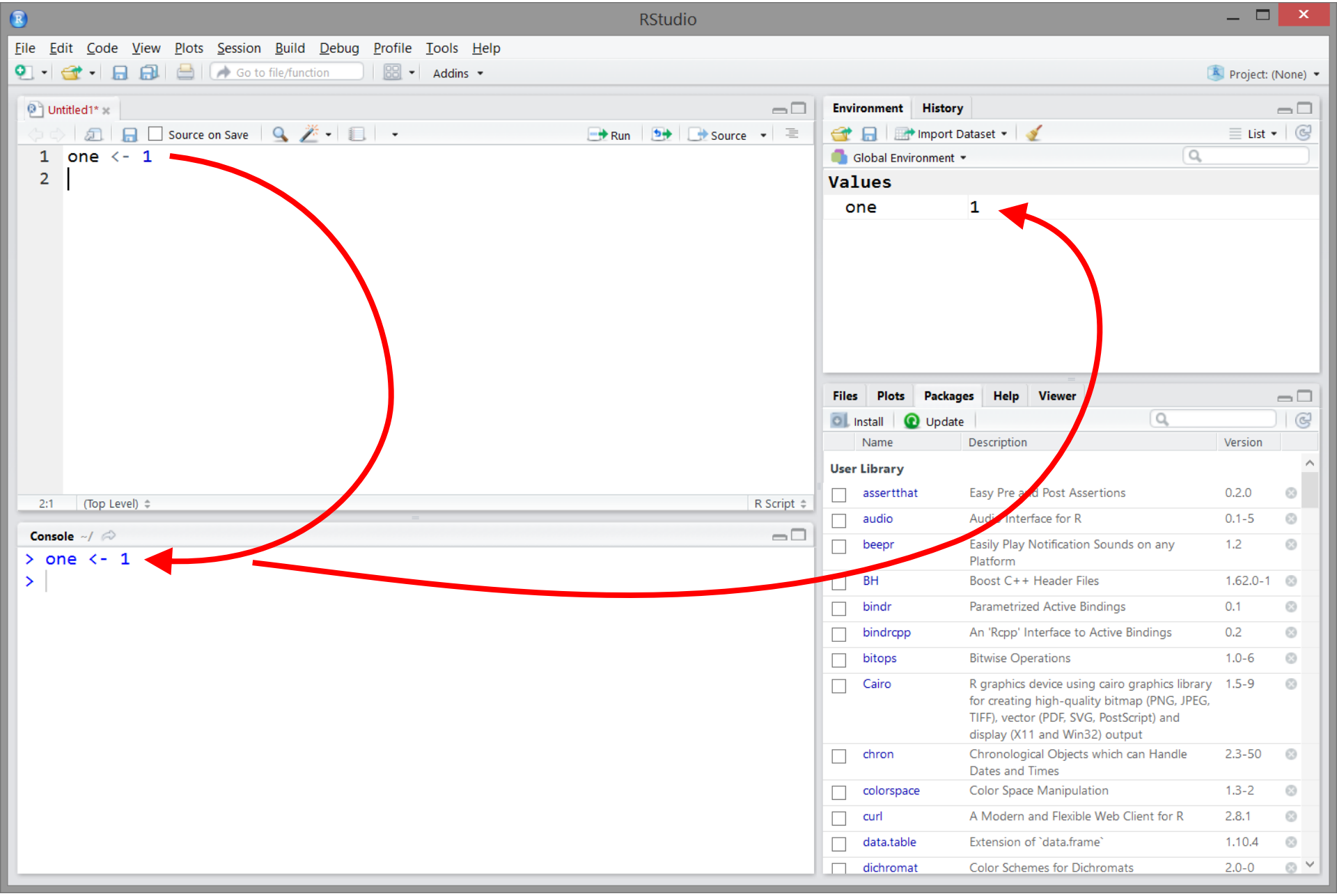
Object

- Object is a container which holds data, and can be manipulated with functions
- The most basic object is called **vector**
- There are other types of objects – **matrix**, **data frame**, **list**

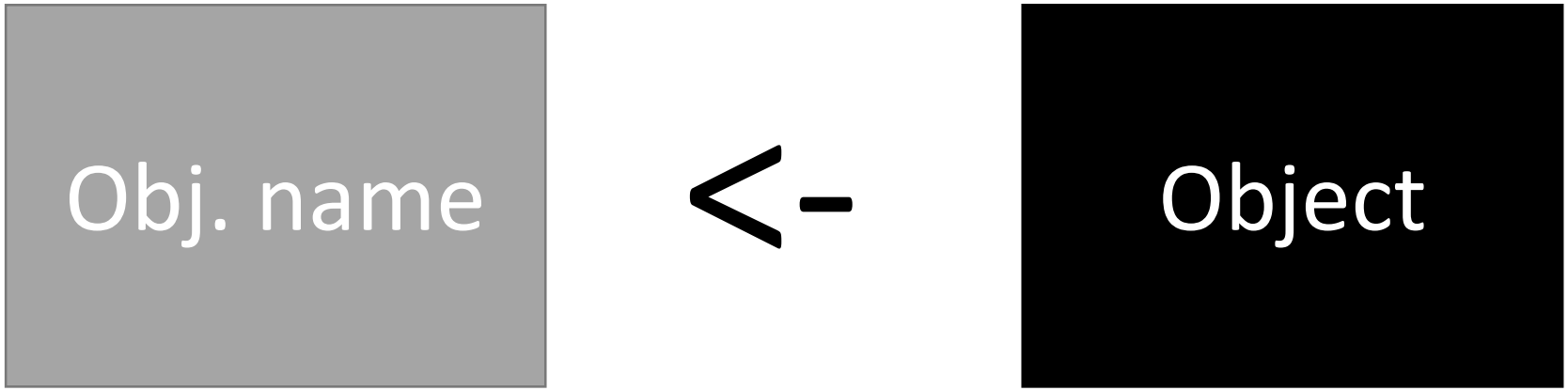
```
one <- 1
```

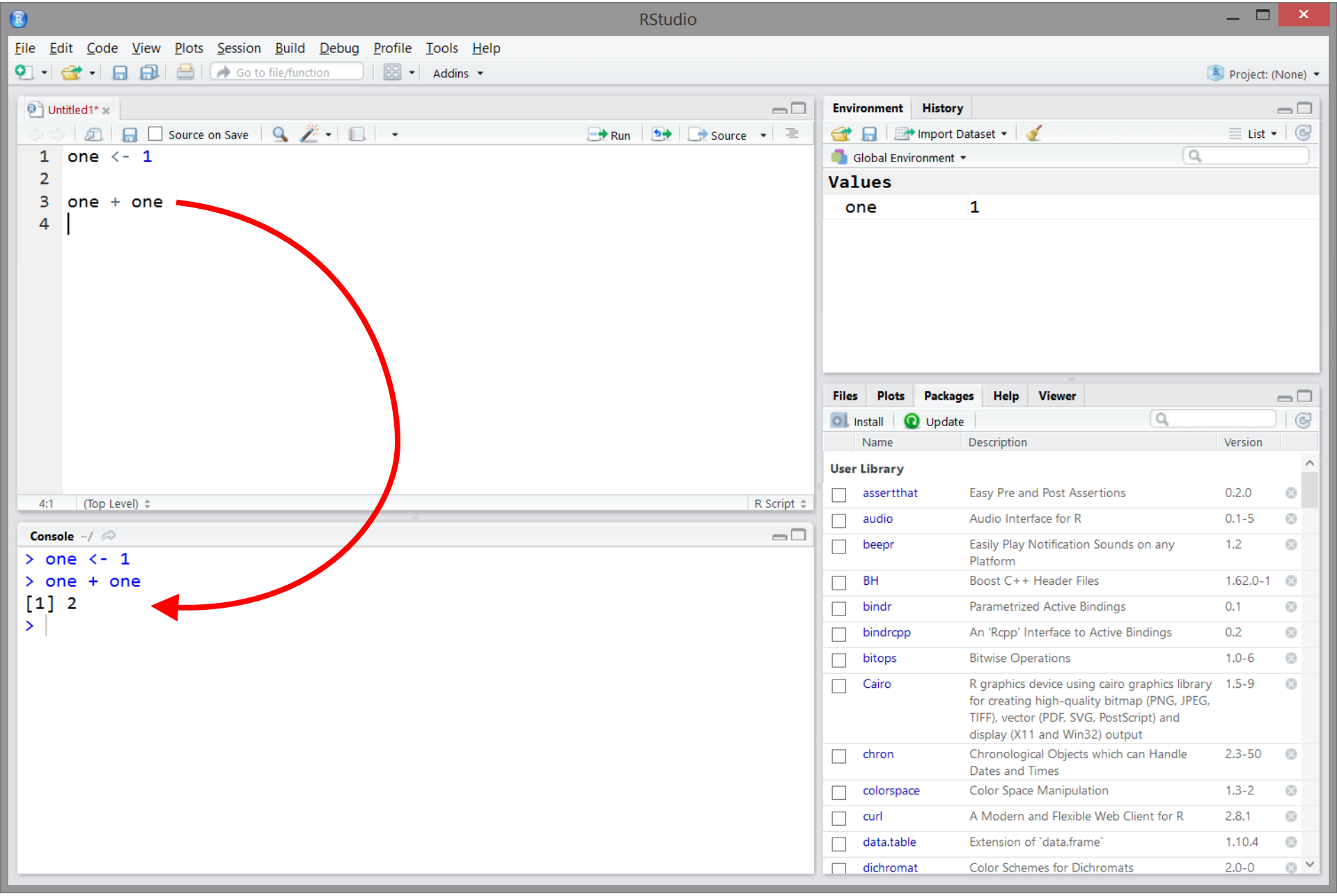

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
1:9
(Top Level)
R Script
Console ~/
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

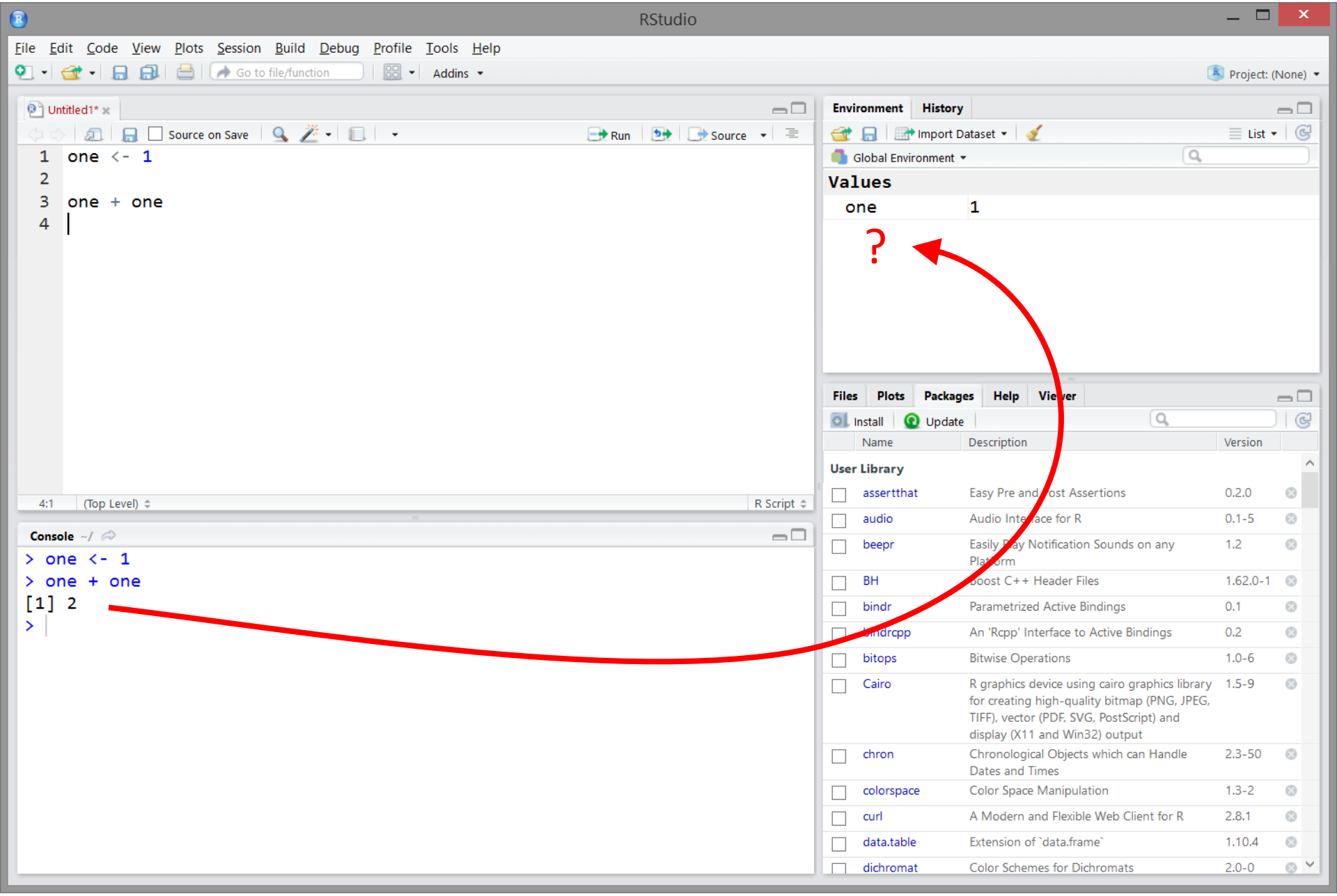

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
1:9
(Top Level)
R Script
Console ~/
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

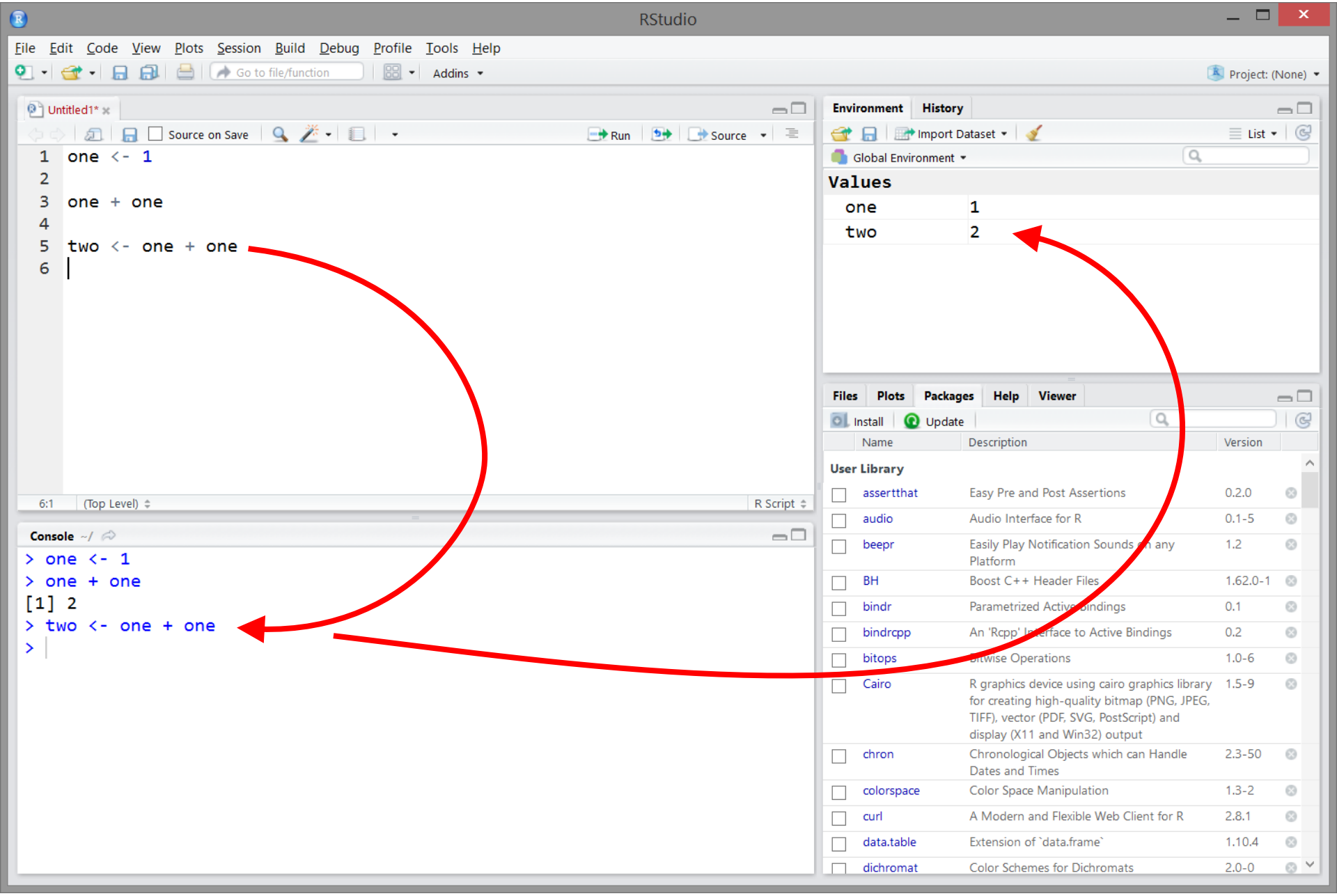

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
2:1 (Top Level)
R Script
Console ~/
> one <- 1
>
Environment History
Import Dataset
List
Global Environment
Values
one 1
Files Plots Packages Help Viewer
Install Update
Name Description Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

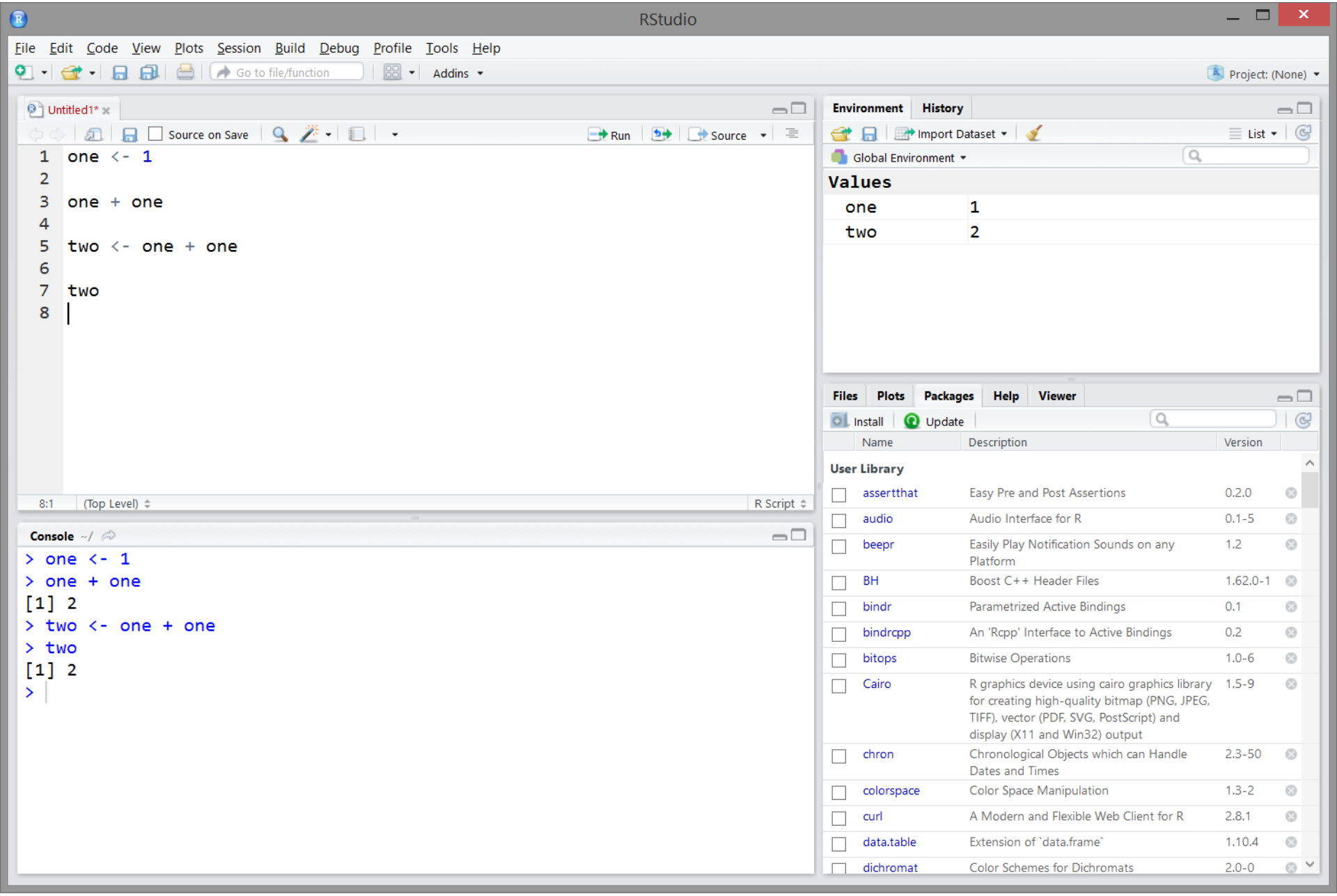
Creating/storing objects




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
4:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
4:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
?
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
5 two <- one + one
6
6:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
> two <- one + one
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
two 2
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
5 two <- one + one
6
7 two
8
8:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
> two <- one + one
> two
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
two 2
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

Functions

- Pre-defined **methods**

- To **create** an object with more than one element, function `c()` is used

```
onetofive <- c(1,3,5,4,2)
```

- Any object may be **manipulated** with a function

```
sort(onetofive)
[1] 1 2 3 4 5
```

Functions

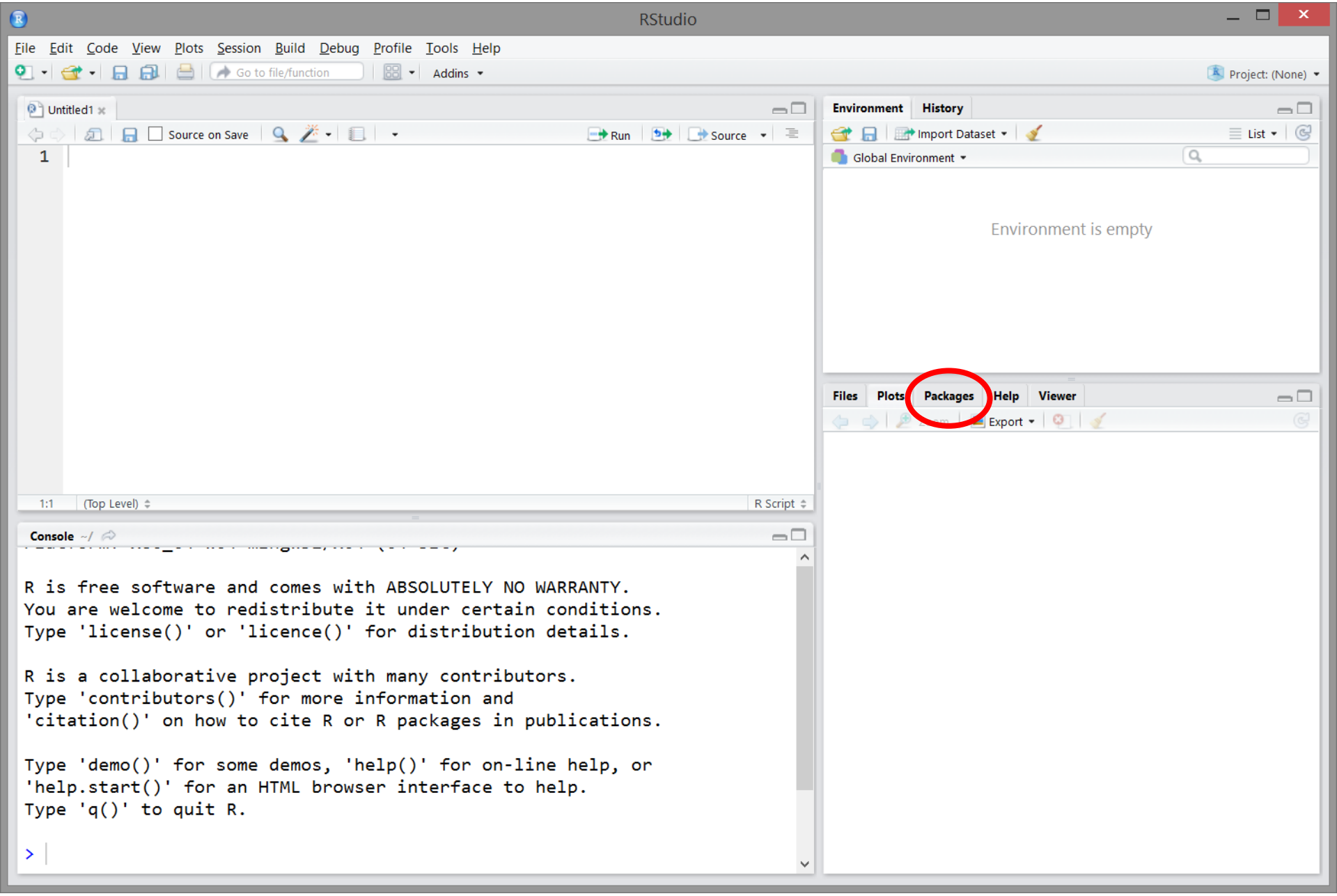
- To extend functionality, functions have pre-defined **arguments**
  - **Arguments are further options** of functions
  - Some functions have many arguments, some none
- To keep function result, it must be stored in the environment as an object

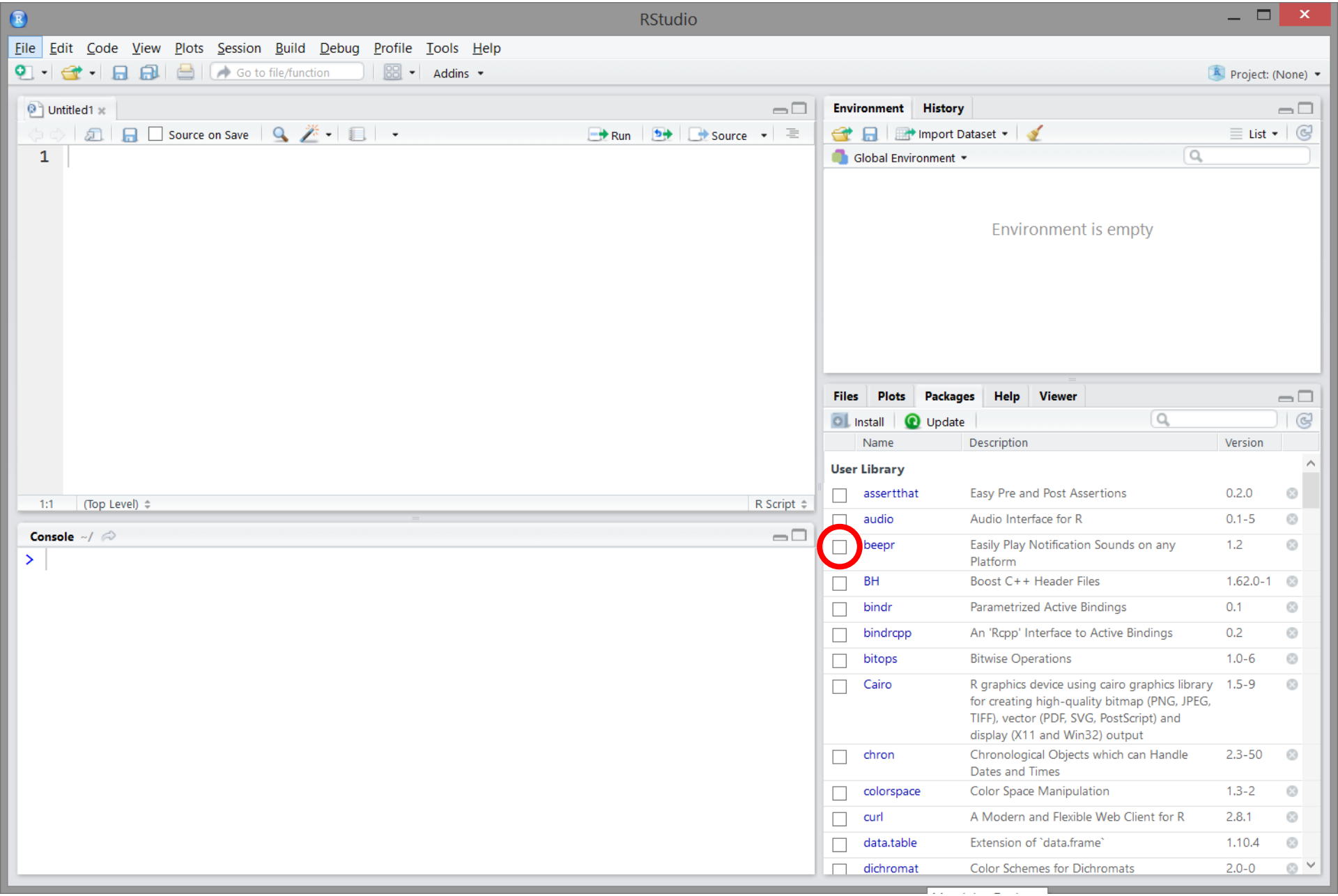
```
sort(onetofive)
[1] 1 2 3 4 5

sort(onetofive, decreasing = TRUE)
[1] 5 4 3 2 1

onetofive <- sort(onetofive, decreasing = TRUE)
```

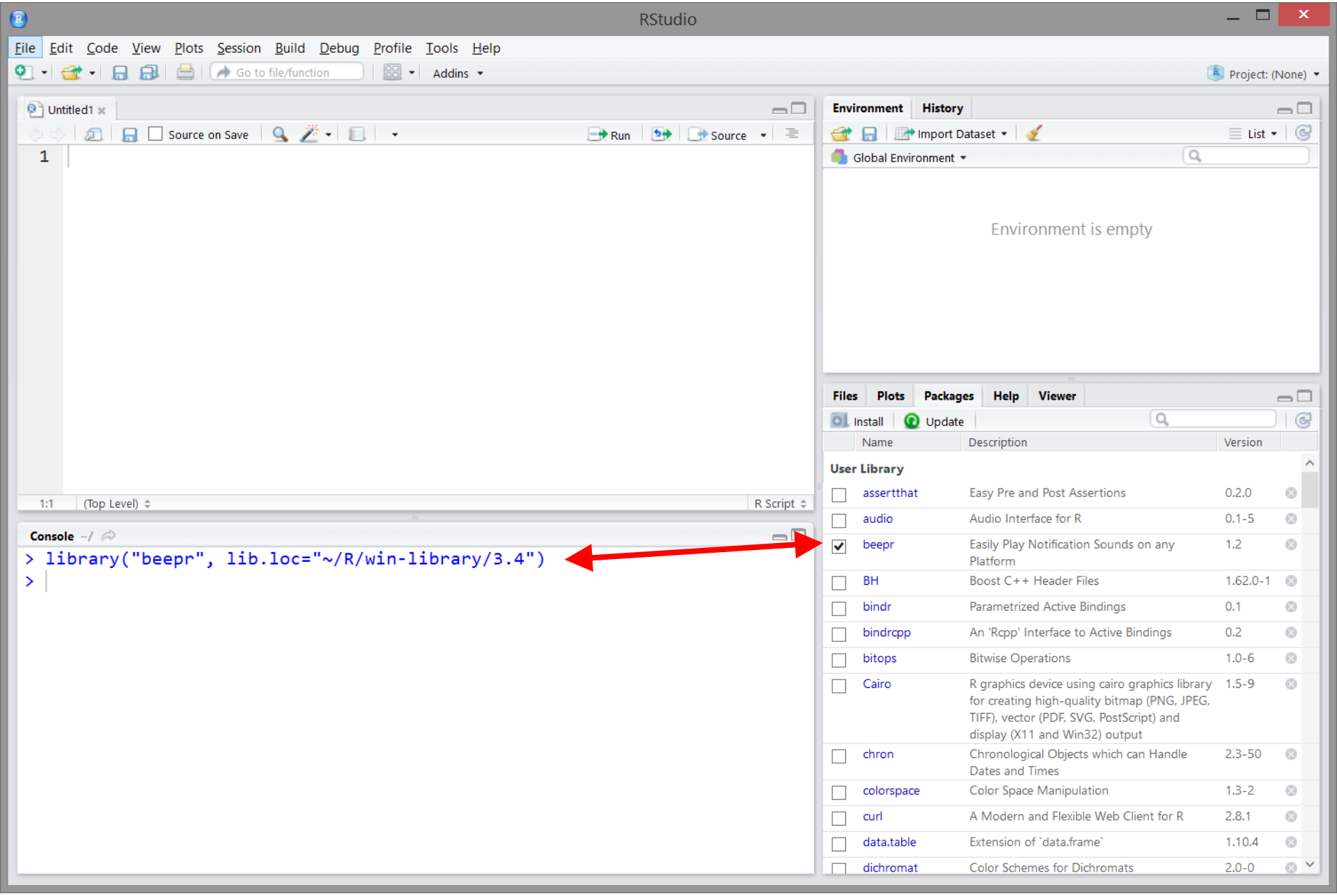
Functions

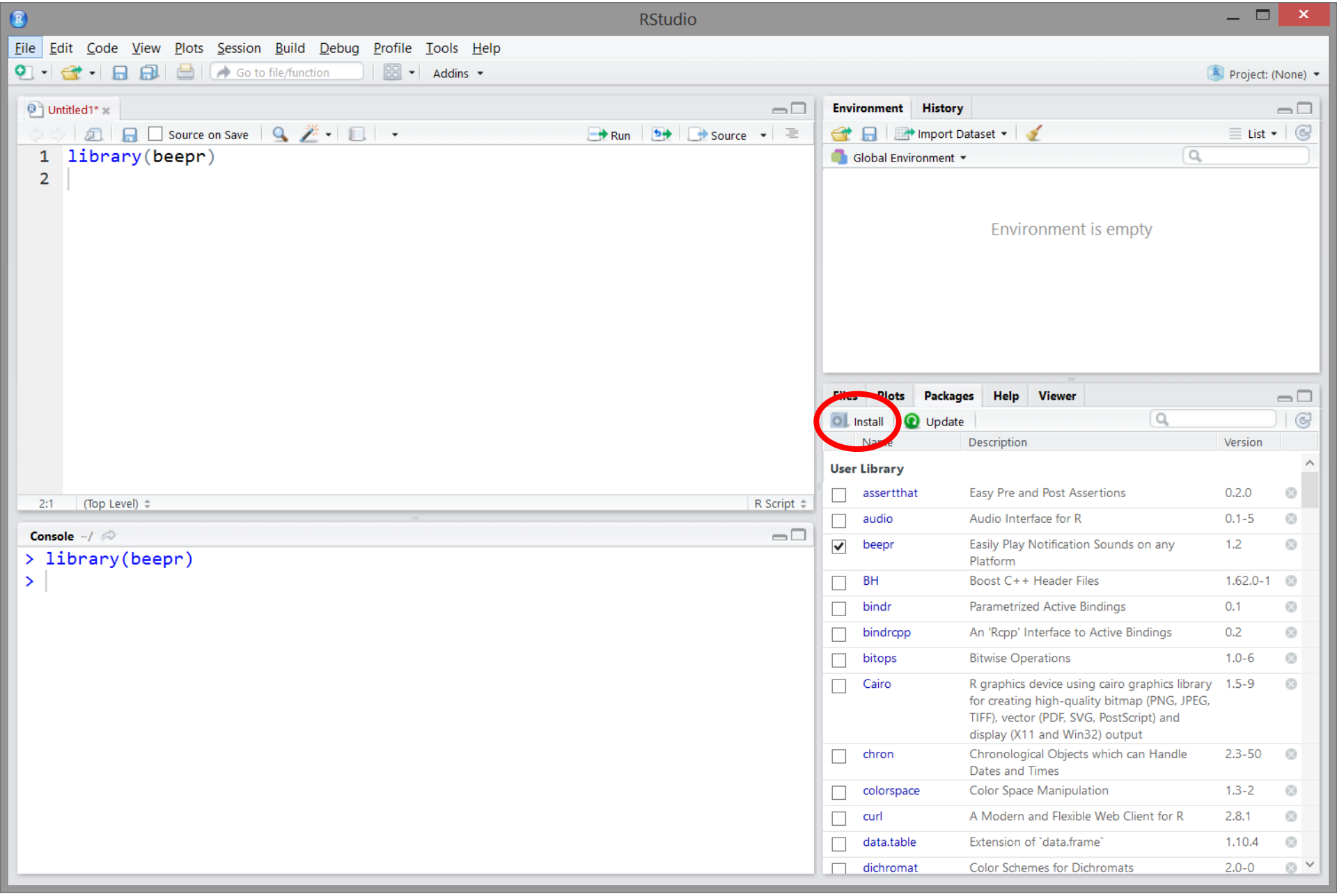
- Arguments usually require input format
  - Boolean input – `TRUE` or `FALSE`
  - Name of object – `onetofive`
  - Text value – `"linear"`
- Format of each argument may be found in **help page**
  - Just **add question mark** in front of the function name

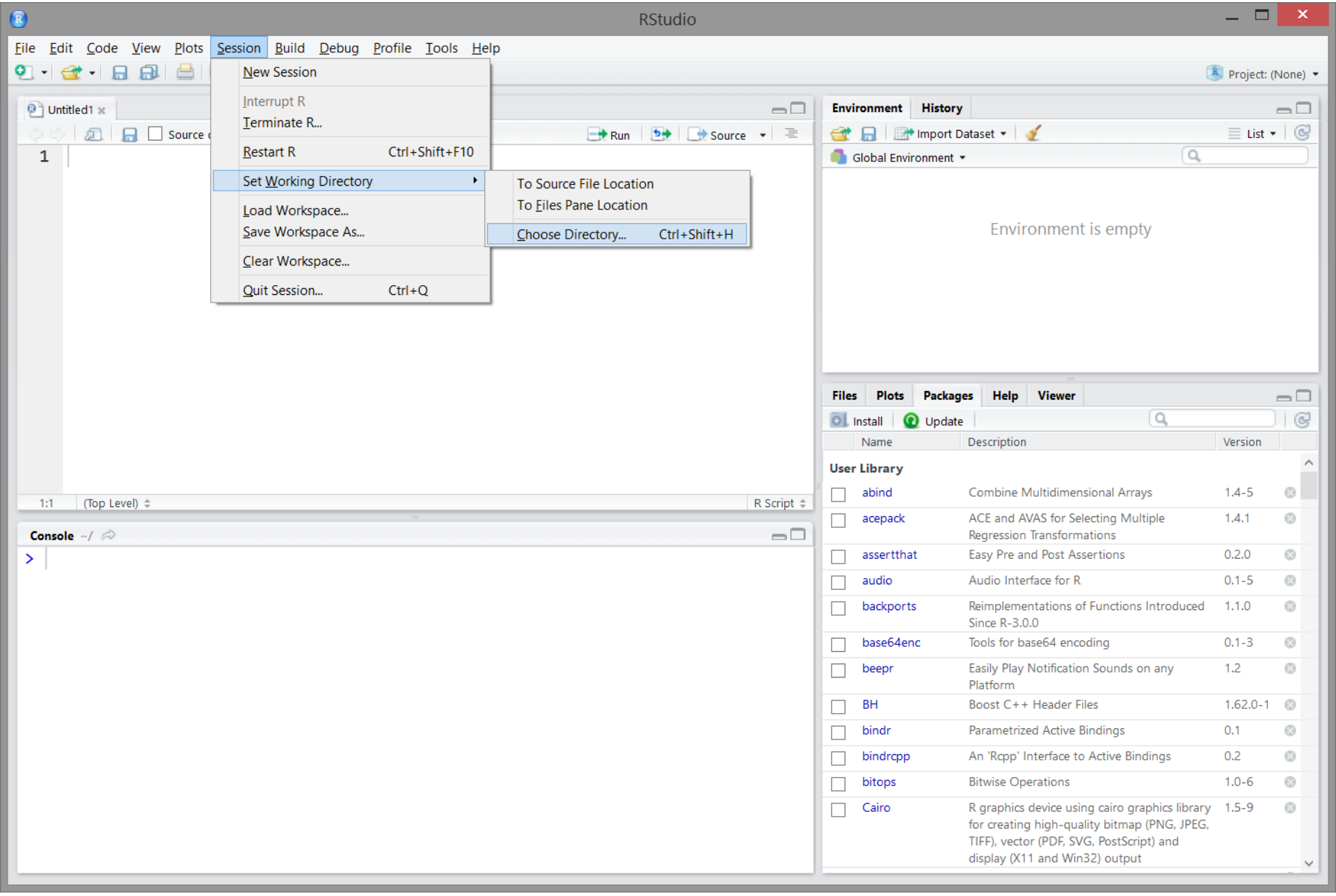
```
?sample()
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
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.
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.
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
> library("beepr", lib.loc="~/R/win-library/3.4")
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
R Script
Console ~/
> library(beepr)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name Description Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

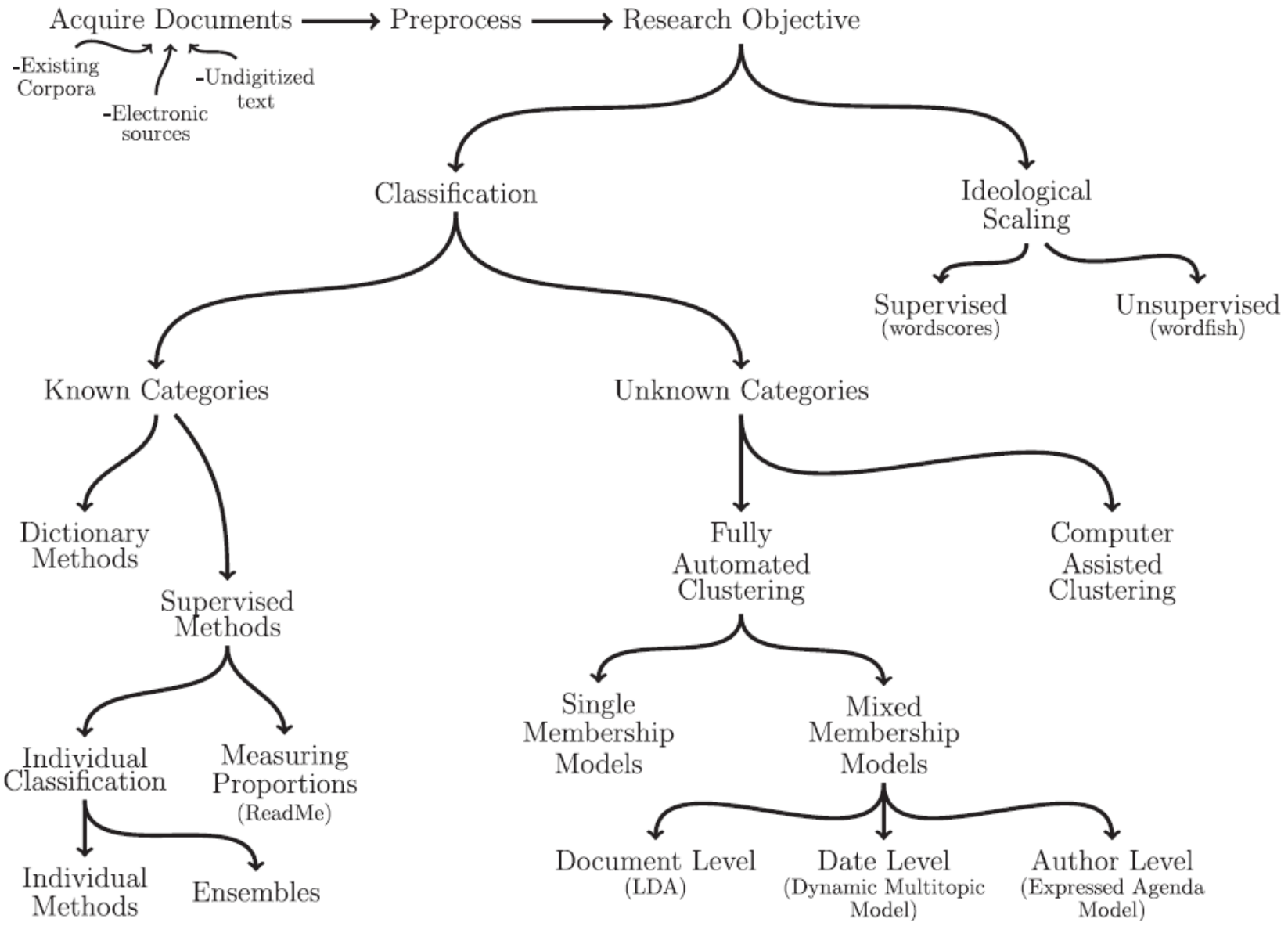

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
New Session
Interrupt R
Terminate R...
Restart R Ctrl+Shift+F10
Set Working Directory
Load Workspace...
Save Workspace As...
Clear Workspace...
Quit Session... Ctrl+Q
To Source File Location
To Files Pane Location
Choose Directory... Ctrl+Shift+H
Project: (None)
Untitled1
Source
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
abind Combine Multidimensional Arrays 1.4-5
acepack ACE and AVAS for Selecting Multiple Regression Transformations 1.4.1
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.0
base64enc Tools for base64 encoding 0.1-3
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9

Data export: saving XLSX

- Package “xlsx”
- Function `write.xlsx()`
- Arguments
  - `x` – object from the environment which you want to export
  - `file` – name of the file in your working directory

```
write.xlsx(x = object, file = "mysheet.xlsx")
```

Quantitative TA in R



(Grimmer & Steweart 2013)

„Be careful what is a result and what is just a residue of your data choices“

Jana Diesner, 2018

Text analysis in R

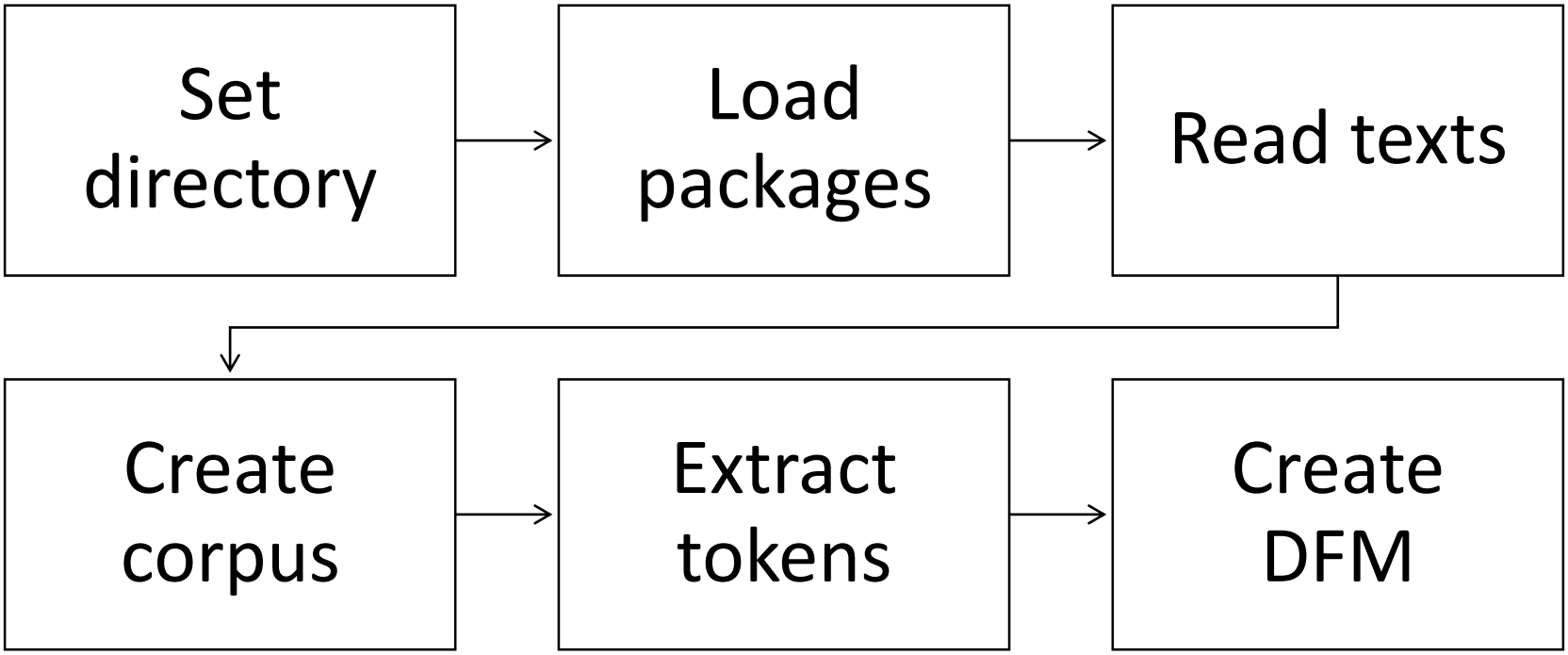
- Package “quanteda” (http://quanteda.io)
  - Developed by Ken Benoit (LSE)
  - Comprehensive package on text analysis methods
- Package “readtext”
  - Ken Benoit & Adam Obeng
  - Package which allows data import from text sources
  - Easy to work with
- Package “stopwords”
  - Ken Benoit, David Muhr & Kohei Watanabe
  - Package containing various stopwords for different languages
- …

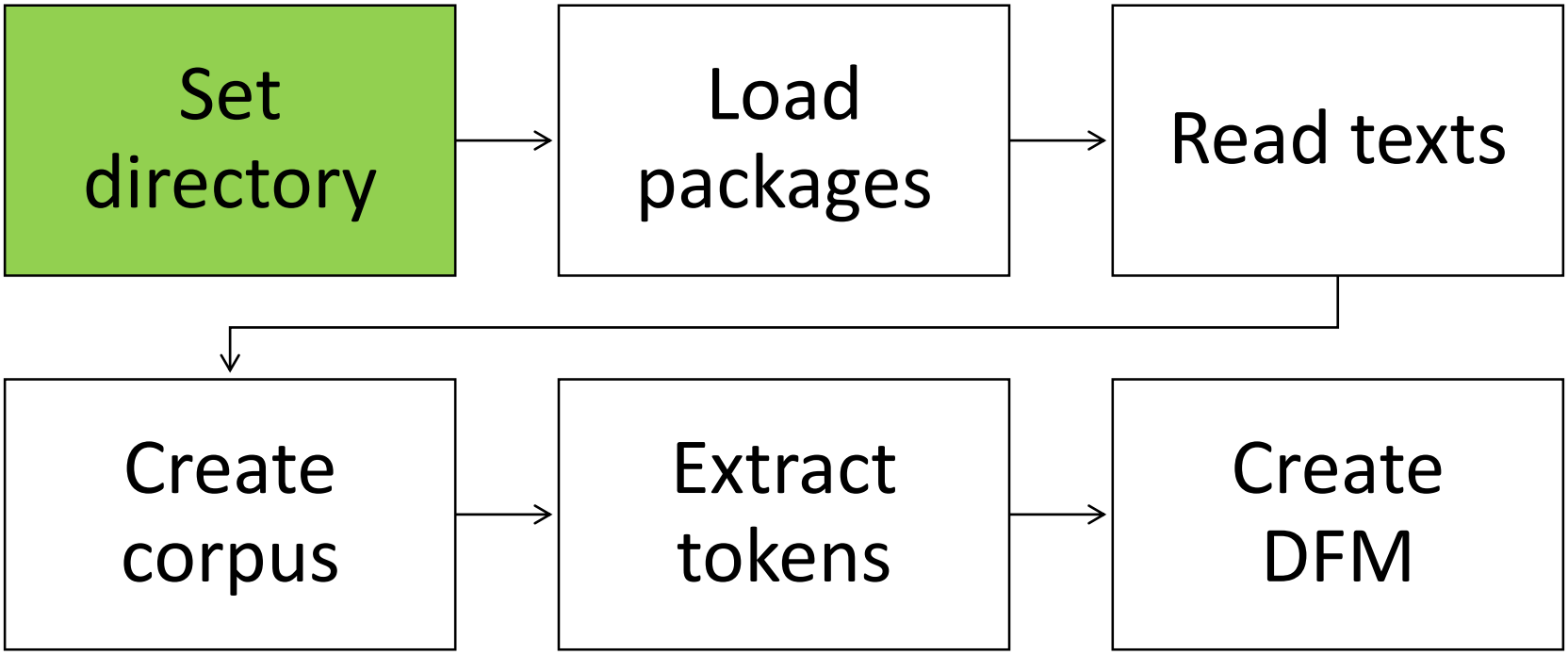
Before we start …

- Open the folder “text_analysis_quanti” folder
- Open script file “text_analysis_1.R” in R Studio

- Install all libraries
  - quanteda, readtext, stopwords, xlsx

Steps leading to analysis



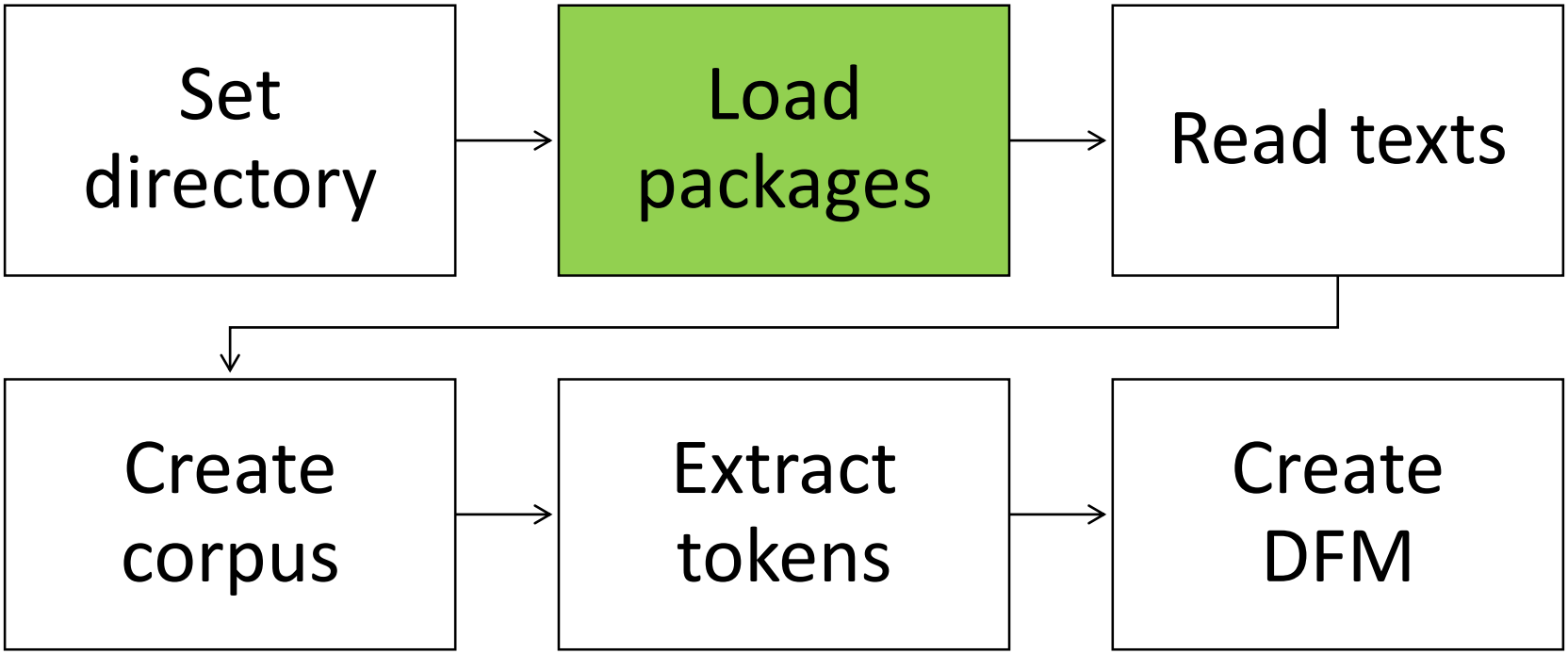

Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

Set working directory

- Window approach
  - Session -> Set Working Directory -> Choose Folder

- Script approach

```
work.dir <- "C:\\path\\to\\folder\\"

setwd(work.dir)
```

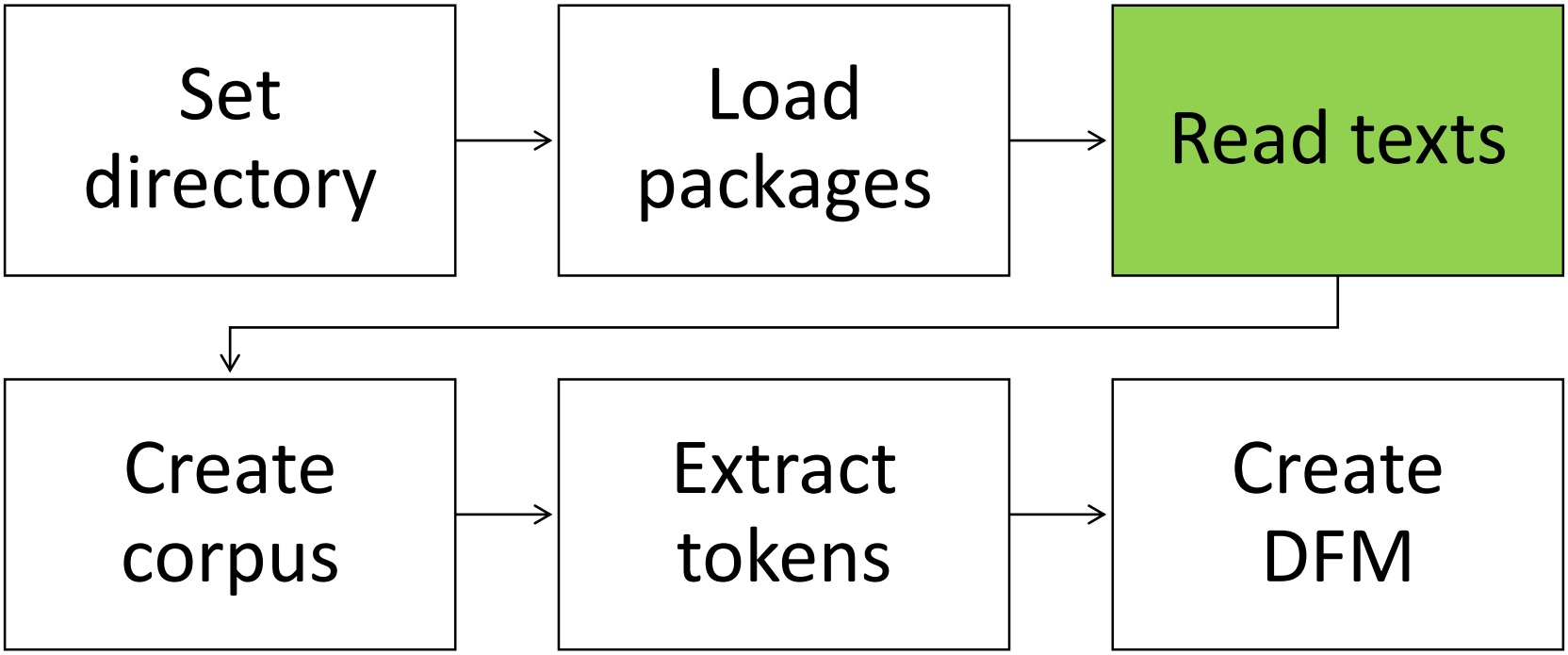

Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

Load packages

- Window approach
  - Session -> Set Working Directory -> Choose Folder

- Script approach

```
library(readtext)
library(quanteda)
library(stopwords)
library(xlsx)
```


Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

Reading texts into R

- `readtext()` function loads **all text files** into R
  - Very **easy to use** – reads everything in any specified folder
  - Supports various document types
    - TXT
    - PDF
    - DOC
    - Twitter data format JSON
    - …
  - Just need to **insert a path to a specific folder**
- Arguments
  - `file`
    - Path to specific source file or path to folder containing files
  - `encoding`

Reading texts into R

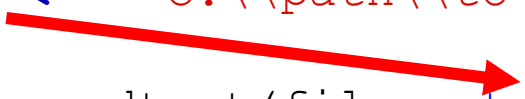
- Encoding
  - Text files are usually stored in certain computer-readable format
- Consider text "*Príklad zlého kódovania*"
  - ASCII/ISO-8859-1: "*PrÃklad zlÃ©ho kÃ³dovania*"
  - UTF-8: "*Príklad zlého kódovania*"
- As a rule of thumb, **UTF-8 encoding** is **desired**

Reading texts into R

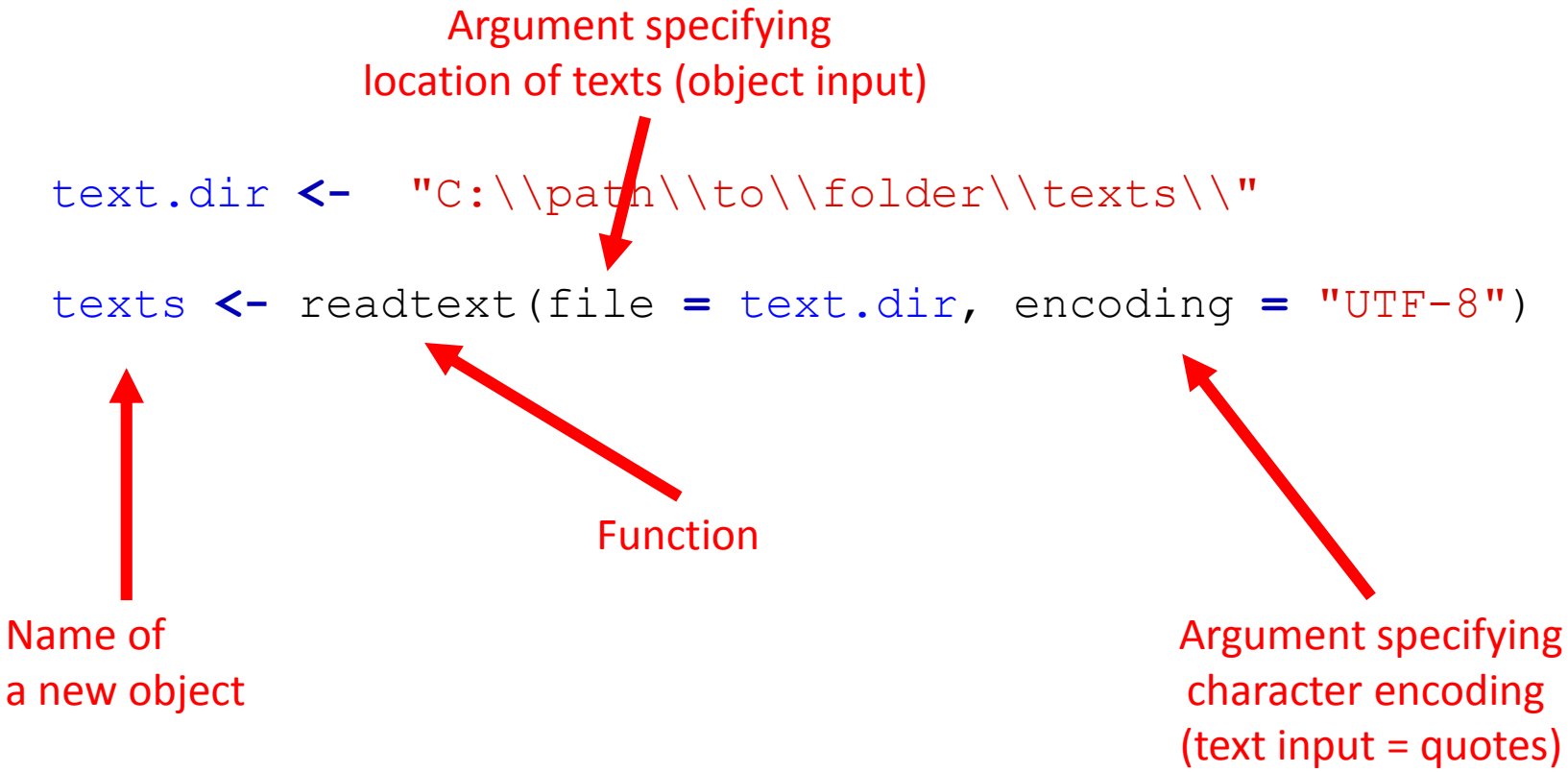
```
text.dir <-  "C:\\path\\to\\folder\\with\\texts\\"

texts <- readtext(file = text.dir, encoding = "UTF-8")
```

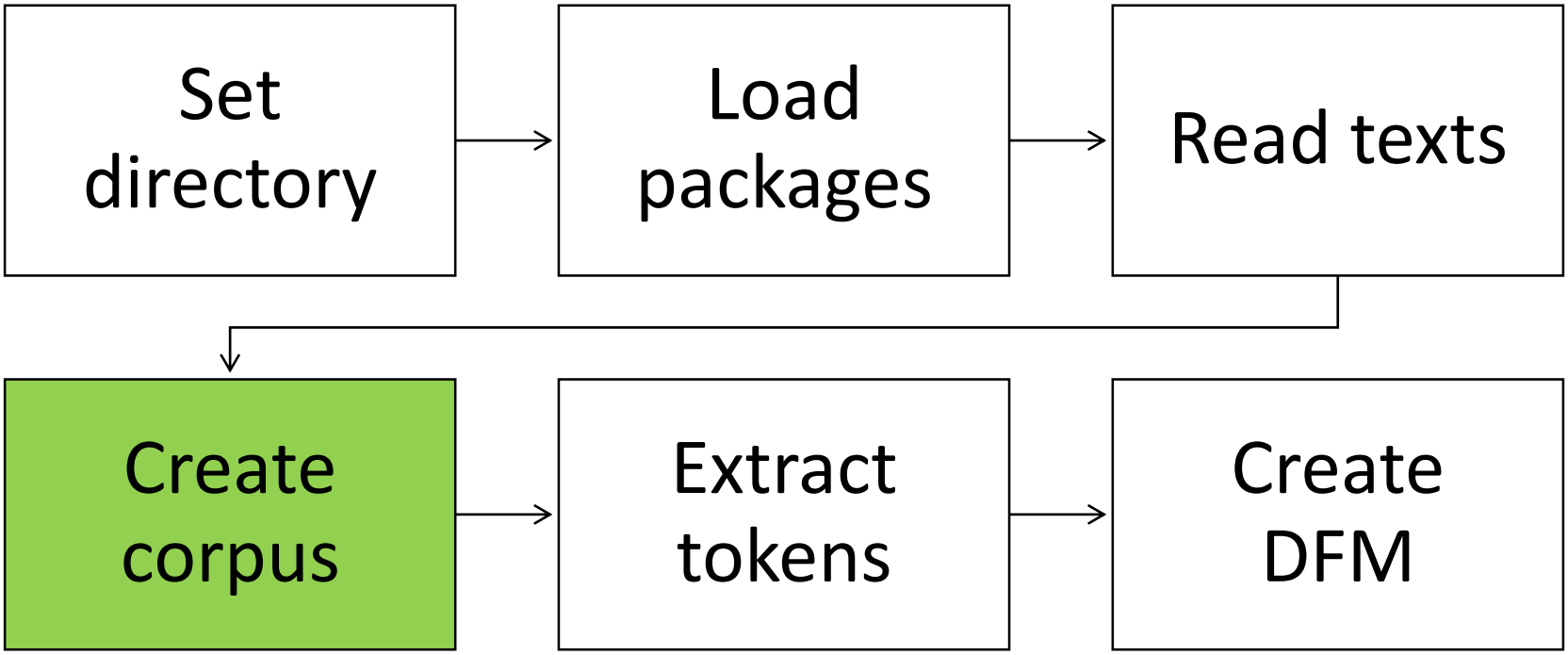
Reading texts into R

```
text.dir <-  "C:\\path\\to\\folder\\texts\\"

texts <- readtext(file = text.dir, encoding = "UTF-8")
```



Reading texts into R




Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

Corpus

- Simple function `corpus()`
  - Creates corpus from all imported texts from the previous step
- Arguments
  - `x`
    - Imported text files
  - `docnames`
    - Optional specification of document names

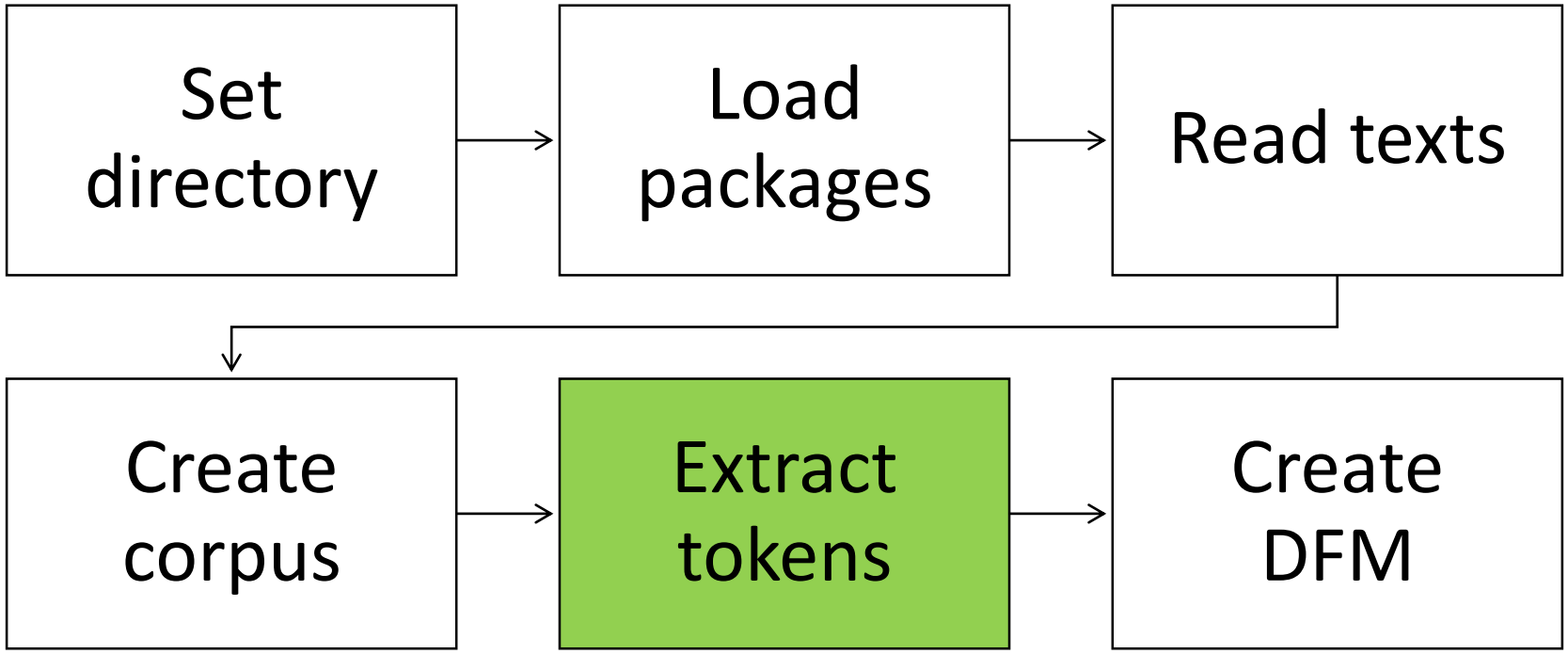
Corpus

- All sorts of statistics may be acquired once corpus is generated

```
corp <- corpus(x = texts)
```

- `summary()`
  - Provides overview of corpus documents
- `ndoc()`
  - Counts number of documents in the corpus

```
ndoc(corp)

summary(corp)
```


Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

From corpus to DFM

- Two-step process
- Tokenization of corpus
  - A step necessary to apply some pre-processing choices which are not text-based (removal of **noise**)
    - Remove numbers
    - Remove punctuation
    - Remove white space (separators)
- DFM generation from tokens
  - Furthter **pre-processing** choices (because of bag-of-words)
    - Stemming
    - Lowercasing
    - Stopwords removal
  - **Dictionary** application

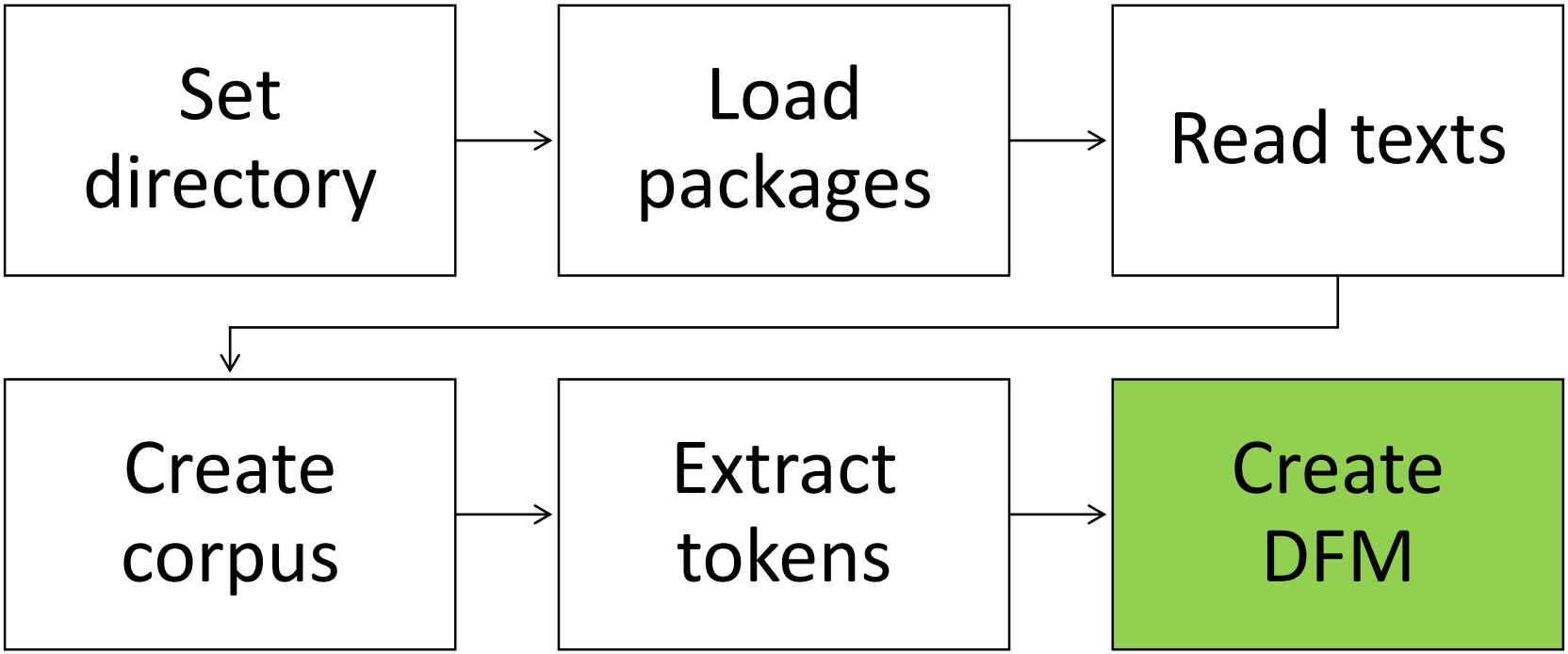
Tokenization

- Function `tokens()`
- Tokenization **arguments**
  - `what` – word, character, sentence
  - `ngrams` – ngramization of the corpus
- Pre-processing **arguments**
  - `remove_numbers` – numerals
  - `remove_punct` – punctuation
  - `remove_symbols` – special “Unicode” symbols (encoding residues)
  - `remove_separators` – white space, line ends, etc.
  - `remove_hyphens` – remove hyphens between words

Tokenization

```
tokenization <- tokens(x = corp,
                       what = "word",
                       ngrams = 1,
                       remove_numbers = TRUE,
                       remove_punct = TRUE,
                       remove_separators = TRUE,
                       remove_hyphens = FALSE )
```

```
tokenization.bigrams <- tokens(x = corp,
                               what = "word",
                               ngrams = c(1:2),
                               remove_numbers = TRUE,
                               remove_punct = TRUE,
                               remove_separators = TRUE,
                               remove_hyphens = FALSE )
```


Set directory
Load packages
Read texts
Create corpus
Extract tokens
Create DFM

Document-feature matrix

- Function `dfm()`
  - Documents in rows, features (tokens) in columns
- Preprocessing **arguments**
  - `tolower` – converts words to lowercase
  - `stem` – implement stemmer
  - `remove` – list of words to be dropped from the DFM
- Application **arguments**
  - `dictionary` – applies dictionary and converts features from tokens to dictionary dimensions
  - `groups` – allows to add another dimension by which the corpus can be grouped/split

Document-feature matrix

```
basic.matrix <- dfm(x = tokenization,
                    tolower = TRUE)


bigram.matrix <- dfm(x = tokenization.bigrams,
                     tolower = TRUE)

stem.matrix <- dfm(x = tokenization,
                   tolower = TRUE,
                   stem = TRUE)

prep.matrix <- dfm(x = tokenization,
                   tolower = TRUE,
                   stem = TRUE,
                   remove = stopwords(language = "en"))
```

DFM weighting

- DFM frequencies are displayed in absolute numbers
  - Document size bias
- `dfm_weight()`
  - Weighting of terms according to document size or other rules
  - Useful to offset the effect of the document size
- `dfm_tfidf()`
  - Incidence of term in document divided by number of documents in which it occurs
  - Useful to find **term importance within document**

DFM weighting

```
weight.matrix <- dfm_weight(prep.matrix,
                            scheme = "prop")

tfidf.matrix <- dfm_tfidf(prep.matrix)
```

DFM manipulation

- `dfm_trim()`
  - Reduction in the dimensionality – removal of very sparse words, very frequent words, etc.
- `dfm_subset()`
  - Subsetting of the the DFM – extraction of DFM portion
- `dfm_sample()`
  - Random sampling from the DFM
  - Useful in various computation-intensive tests

DFM manipulation

```
red.matrix <- dfm_trim(prep.matrix,
                       min_termfreq = 5)

sample.matrix <- dfm_sample(prep.matrix,
                            size = 5)
```

Analysis

Analysis

- Corpus-based
  - Require full texts
  - E.g. KWIC
- DFM-based
  - Require frequencies
  - Bag-of-words assumption
  - E.g. token frequencies, correspondence analysis, wordfish …

Keywords in context

- `kwic()` function
- Modifying arguments
  - `pattern`
    - A term of interest or multiple terms of interest wrapped in function `c()`
  - `window`
    - Length of text part extracted before and after the keyword term
  - `case.insensitive`
    - Binary – should function take term case into account?
- You can save it using `write.xlsx()` function

Keywords in context

```
kewords.in.context <- kwic(corp,
                           pattern = "energy",
                           window = 5)


kewords.in.context.2 <- kwic(corp,
                             pattern = c("energy", "russia"),
                             window = 5)

write.xlsx(x = kewords.in.context,file = "kwic.xlsx")
```

Keywords in context

Docname	From	To	Pre	Keyword	Post
2007-2008-CZ.txt.proc.txt	3784	3784	current issues related to renewable	**energy**	sources , cooperation in EU
2007-2008-CZ.txt.proc.txt	3805	3805	The V4 working group on	**energy**	meets regularly . The European
2007-2008-CZ.txt.proc.txt	3842	3842	cooperation in the field of	**energy**	with Nordic Council countries .
2008-2009-PL.txt.proc.txt	101	101	in early 2009 the Russian-Ukrainian	**energy**	crisis broke out , with
2008-2009-PL.txt.proc.txt	1195	1195	progress , the issue of	**energy**	security became the prime topic
2008-2009-PL.txt.proc.txt	1269	1269	group of governmental plenipotentiaries for	**energy**	security . On June 3rd
2008-2009-PL.txt.proc.txt	1580	1580	. September 5th 2008 -	**Energy**	Expert Group meeting . The
2008-2009-PL.txt.proc.txt	1613	1613	Agency for the Co-operation of	**Energy**	Regulations [ACER] ,

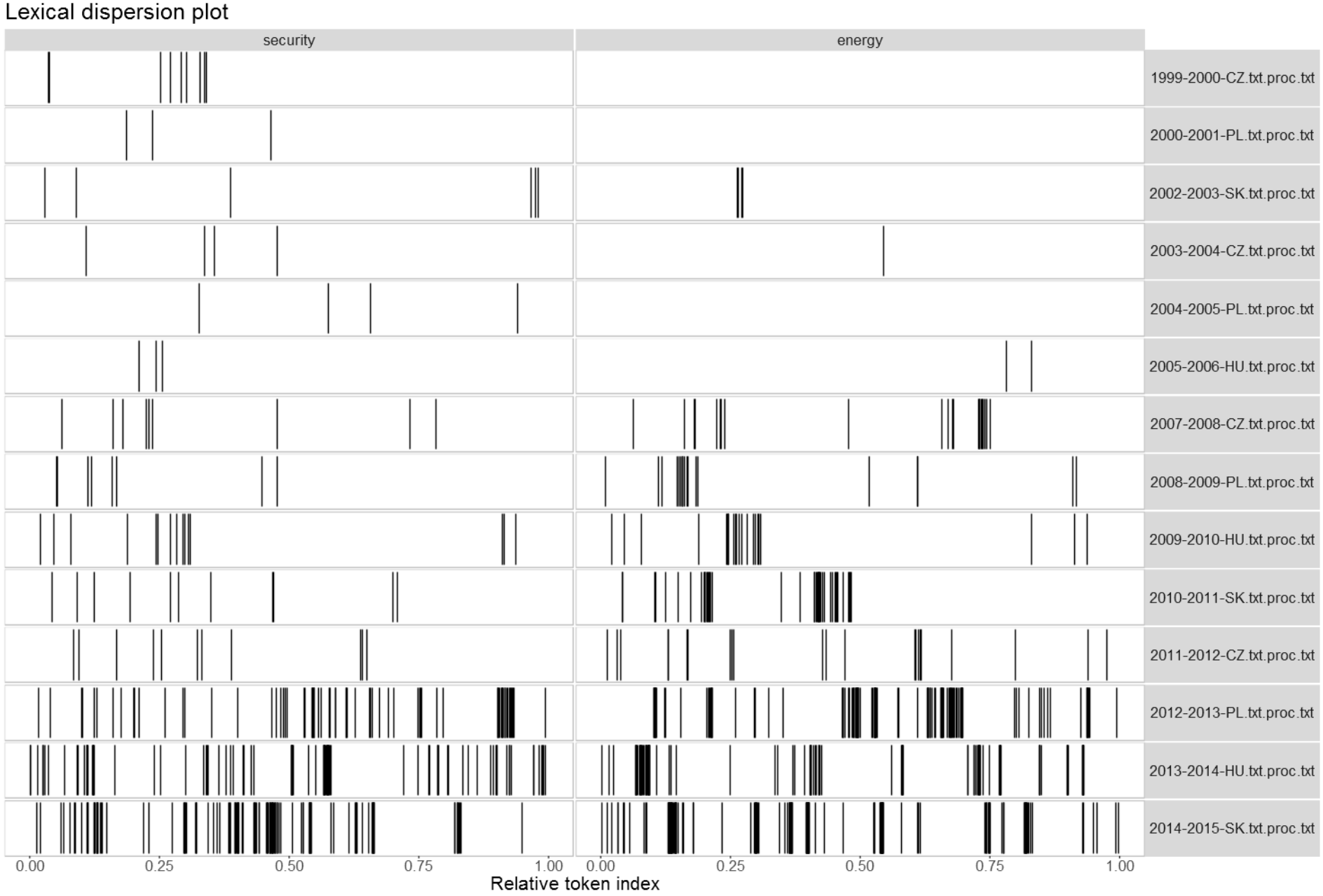
Keywords in context

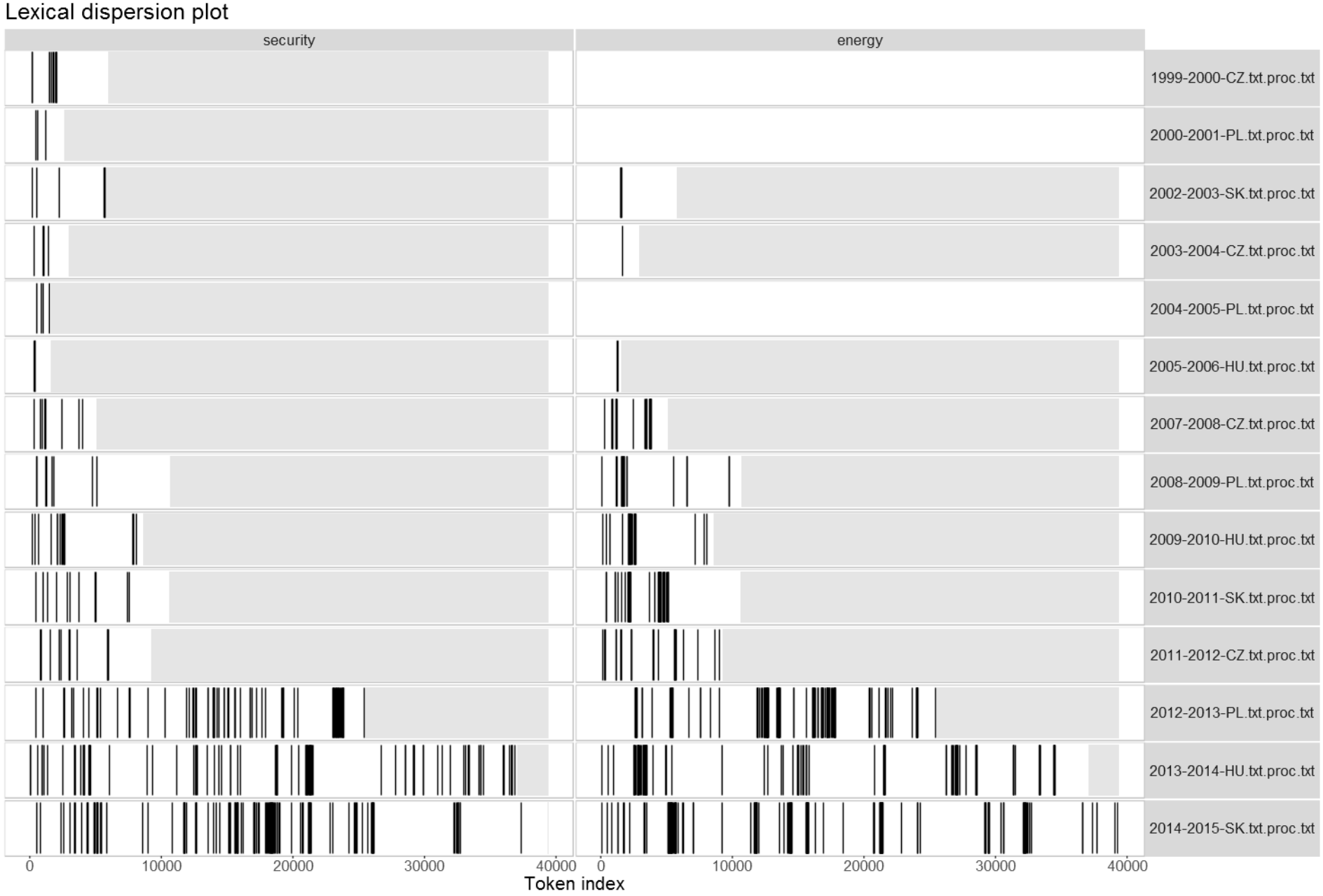
- May be plotted easily
- `textplot_xray()`
  - Function for plotting
  - One or several KWIC objects may be passed (each must be passed separately)
  - Argument `scale` allows to plot absolute or weighted positions (normalized by document length)

Keywords in context

```
textplot_xray(
       kwic(corp, pattern = "energy", window = 5),
       kwic(corp, pattern = "security", window = 5),
       sort = TRUE
)
```

```
 textplot_xray(
        kwic(corp, pattern = "energy", window = 5),
        kwic(corp, pattern = "security", window = 5),
        sort = TRUE,
        scale = "absolute"
 )
```


Lexical dispersion plot
security
energy
1999-2000-CZ.txt.proc.txt
2000-2001-PL.txt.proc.txt
2002-2003-SK.txt.proc.txt
2003-2004-CZ.txt.proc.txt
2004-2005-PL.txt.proc.txt
2005-2006-HU.txt.proc.txt
2007-2008-CZ.txt.proc.txt
2008-2009-PL.txt.proc.txt
2009-2010-HU.txt.proc.txt
2010-2011-SK.txt.proc.txt
2011-2012-CZ.txt.proc.txt
2012-2013-PL.txt.proc.txt
2013-2014-HU.txt.proc.txt
2014-2015-SK.txt.proc.txt
Document
0.00
0.25
0.50
0.75
1.00
Relative token index


Lexical dispersion plot
security
energy
1999-2000-CZ.txt.proc.txt
2000-2001-PL.txt.proc.txt
2002-2003-SK.txt.proc.txt
2003-2004-CZ.txt.proc.txt
2004-2005-PL.txt.proc.txt
2005-2006-HU.txt.proc.txt
2007-2008-CZ.txt.proc.txt
2008-2009-PL.txt.proc.txt
2009-2010-HU.txt.proc.txt
2010-2011-SK.txt.proc.txt
2011-2012-CZ.txt.proc.txt
2012-2013-PL.txt.proc.txt
2013-2014-HU.txt.proc.txt
2014-2015-SK.txt.proc.txt
Document
0
10000
20000
30000
40000
Token index

Frequencies

- Frequency of **features** in the DFM
  - Absolute token frequencies
  - Dictionary category frequencies
- `topfeatures()`
  - General function to extract number of tokens

Frequencies

```
freq.basic <- textstat_frequency(x = basic.matrix, n = 20)

freq.stem <- textstat_frequency(x = stem.matrix, n = 20)

freq.prep <- textstat_frequency(x = prep.matrix, n = 20)

write.xlsx(x = freq.prep, file = "frequencies.xlsx")
```

Wordcloud

- Function `textplot_wordcloud()`

Argument	Description
`x`	Terms
`max_words`	Maximum number of words rendered
`min_size`	Size of smallest category
`max_size`	Size of largest category
`rotation`	Percentage of terms placed vertically
`color`	Color or color palette
`...`	Many other arguments available (use help)

Wordcloud

```
textplot_wordcloud(x = basic.matrix,
                   max_words = 50,
                   min_size = 1,
                   max_size = 4,
                   rotation = 0,
                   color = "steelblue2")
```

```
textplot_wordcloud(x = prep.matrix,
                   max_words = 50,
                   min_size = 1,
                   max_size = 4,
                   rotation = 0,
                   color = "red3")
```


regional
security
this
defence
support
be
european
common
at
countries
presidency
slovak
was
in
and
by
czech
we
a
is
visegrad
group
energy
as
the
for
from
its
with
on
an
eastern
joint
v4
of
to
eu
foreign
policy
that
ministers
also
meeting
their
which
cooperation
republic
development
international
budapest


implement
partnership
eastern
exchang
common
market
import
ukrain
issu
affair
relat
european
also
czech
area
develop
minist
integr
ministri
countri
june
group
state
work
intern
europ
presid
v4
eu
joint
new
well
polici
cooper
energi
meet
republ
visegrad
foreign
slovak
region
support
project
expert
secur
discuss
nation
defenc
budapest
activ

Dictionaries

- Two step process
- Requires a dictionary object
  - Manually constructed dictionary
  - Dictionary in the external location
    - File “LaverGarry.cat” in your folder
  - Dictionary included in a package
    - Package “tidytext”
    - Sentiments dictionary
- Dictionary has to be applied in a DFM construction process

Dictionaries

CULTURE

CULTURE-HIGH

ART (1)

ARTISTIC (1)

DANCE (1)

GALLER* (1)

MUSEUM* (1)

MUSIC* (1)

OPERA* (1)

THEATRE* (1)

CULTURE-POPULAR

MEDIA (1)

SPORT

ANGLER* (1)

PEOPLE (1)

WAR_IN_IRAQ (1)

CIVIL_WAR (1)

ECONOMY

+STATE+

ACCOMMODATION (1)

AGE (1)

AMBULANCE (1)

ASSIST (1)

BENEFIT (1)

CARE (1)

CARER* (1)

CHILD* (1)

CLASS (1)

CLASSES (1)

CLINICS (1)

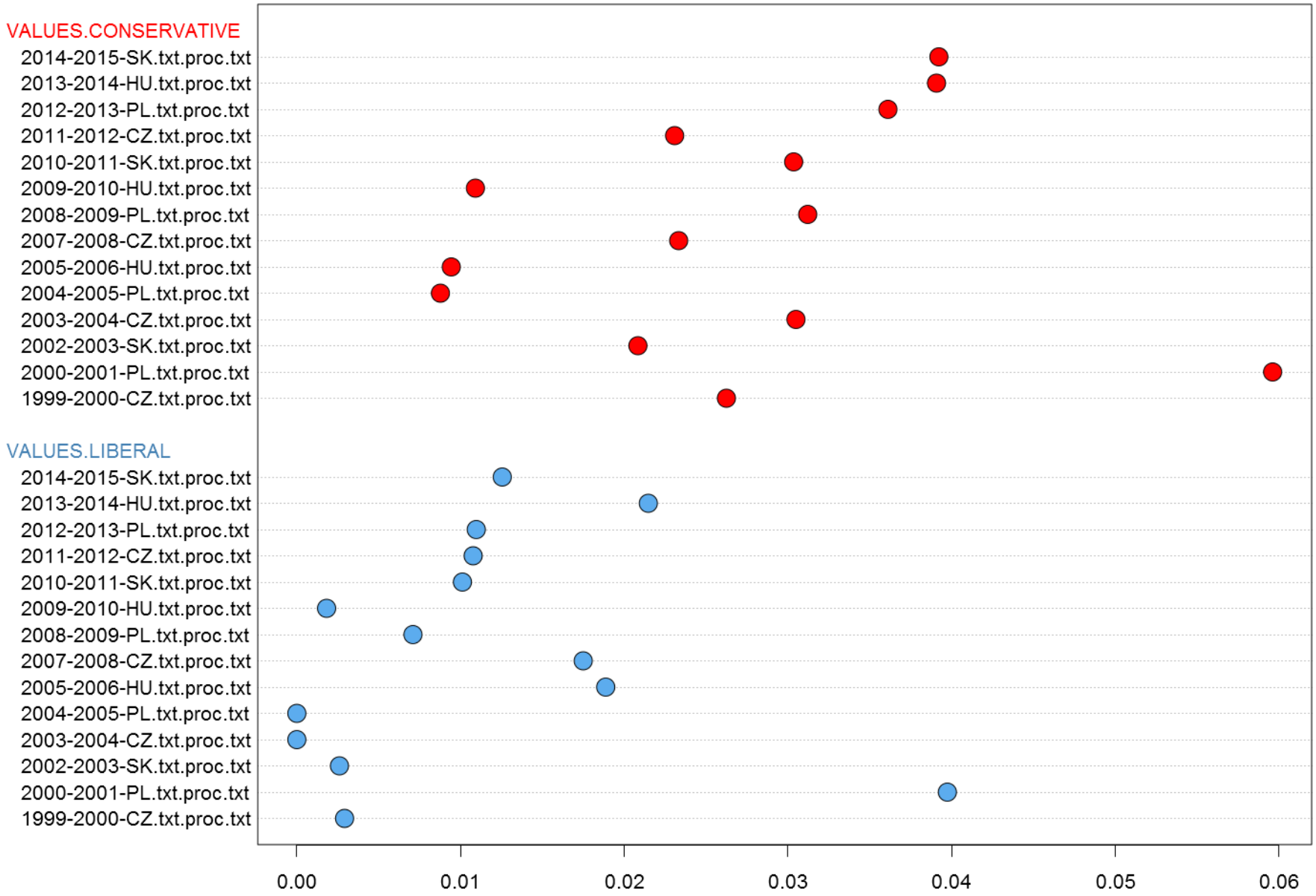
COLLECTIVE* (1)

Dictionaries

- Using dataset from a file/creating own dictionary
  - Function `dictionary()` allows to load a file as a dictionary
  - Arguments
    - `file` – specifies the path to file (because we are in a working directory, we have to specify only a file name)
    - `format` – specifies the pre-defined format of dictionary
- The new object will be used in the DFM argument `dictionary`
- Useful to weigh the DFM after application

Dictionaries

```
wordstat.dict <- dictionary(file = "LaverGarry.cat",
                            format = "wordstat")

dfm.dict <- dfm(tokenization,
                dictionary = wordstat.dict)

dfm.dict.w <- dfm_weight(dfm.dict,scheme = "prop")
```


VALUES.CONSERVATIVE
2014-2015-SK.txt.proc.txt
2013-2014-HU.txt.proc.txt
2012-2013-PL.txt.proc.txt
2011-2012-CZ.txt.proc.txt
2010-2011-SK.txt.proc.txt
2009-2010-HU.txt.proc.txt
2008-2009-PL.txt.proc.txt
2007-2008-CZ.txt.proc.txt
2005-2006-HU.txt.proc.txt
2004-2005-PL.txt.proc.txt
2003-2004-CZ.txt.proc.txt
2002-2003-SK.txt.proc.txt
2000-2001-PL.txt.proc.txt
1999-2000-CZ.txt.proc.txt
VALUES.LIBERAL
2014-2015-SK.txt.proc.txt
2013-2014-HU.txt.proc.txt
2012-2013-PL.txt.proc.txt
2011-2012-CZ.txt.proc.txt
2010-2011-SK.txt.proc.txt
2009-2010-HU.txt.proc.txt
2008-2009-PL.txt.proc.txt
2007-2008-CZ.txt.proc.txt
2005-2006-HU.txt.proc.txt
2004-2005-PL.txt.proc.txt
2003-2004-CZ.txt.proc.txt
2002-2003-SK.txt.proc.txt
2000-2001-PL.txt.proc.txt
1999-2000-CZ.txt.proc.txt
0.00
0.01
0.02
0.03
0.04
0.05
0.06

Second data set

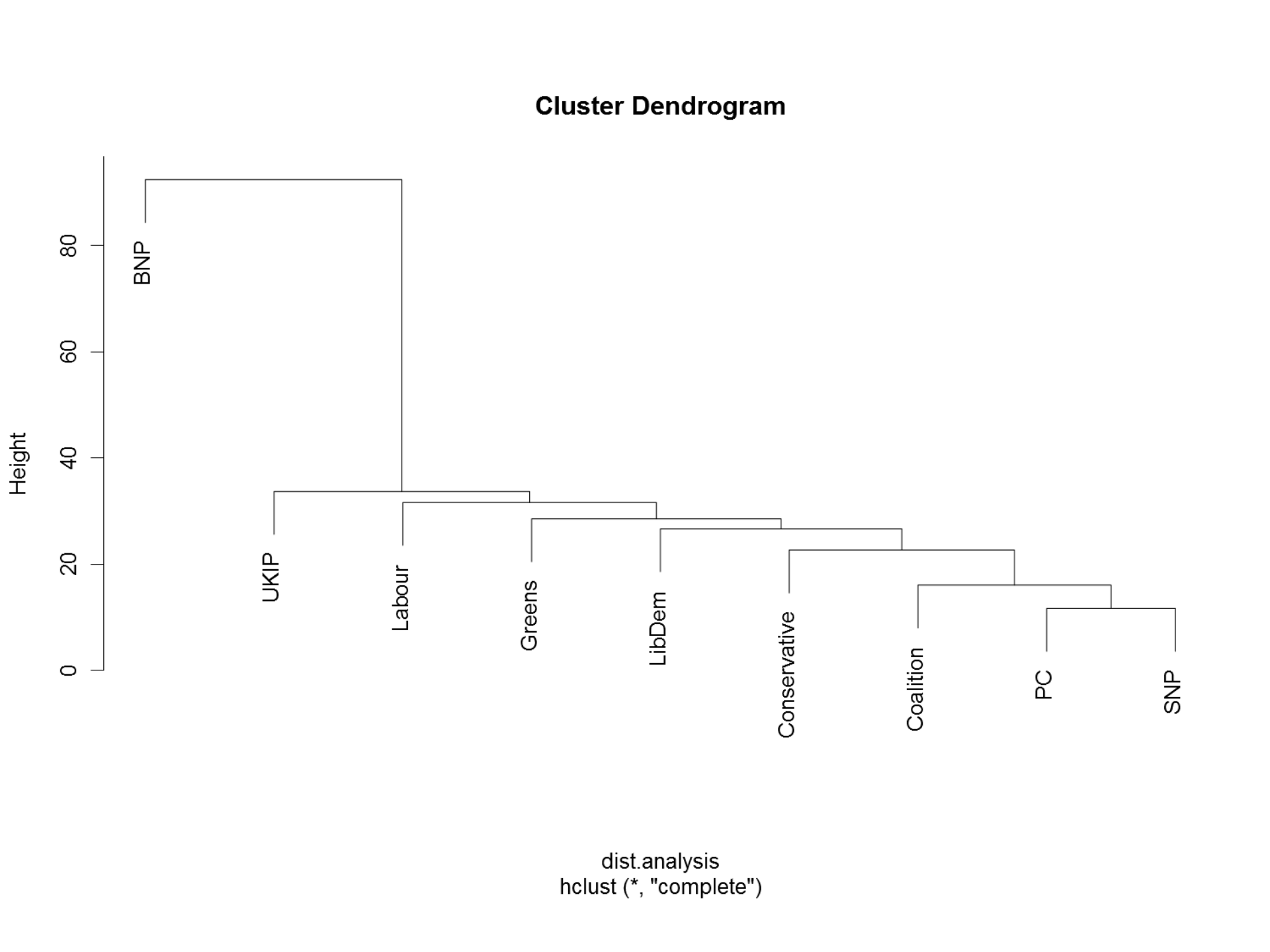
- Parts of UK 2010 election manifestos
  - Issue of migration
  - English, already pre-formatted, part of quanteda package
  - Just type `data_char_ukimmig2010` into script
- Same drill as before
  - Texts
  - Corpus
  - Tokens
  - DFM

Distances

- The simplest algorithm to obtain scaling
- Function `textstat_dist()`
- Creates a distance object which is recognized by other R packages and functions
- We may use `hclust()` function which creates a hierarchical clusters and plot it with `plot()` function afterwards

Distances

```
dist.analysis <- textstat_dist(mig.dfm)

clusters <- hclust(dist.analysis)

plot(clusters)
```

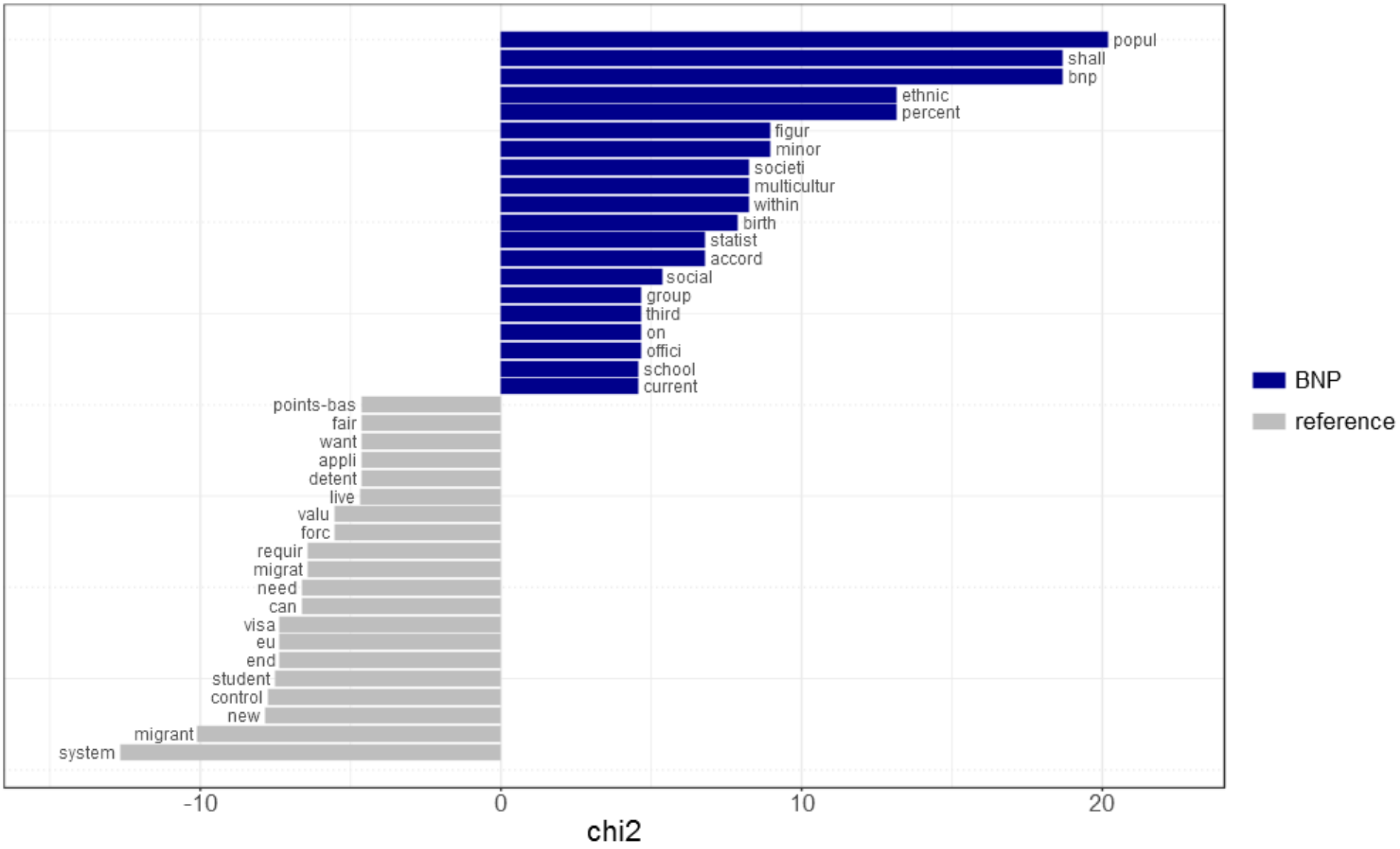

Cluster Dendrogram
Height
0
20
40
60
80
BNP
UKIP
Labour
Greens
LibDem
Conservative
Coalition
PC
SNP
dist.analysis
hclust (*, "complete")

Keyness

- Useful method to evaluate keywords – finds words which are specific in relation to the rest of the corpus
- Function `textstat_keyness()`
  - Argument target
    - Specifies the numeric ID of the document, which is compared to the rest of the corpus
- Can be plotted via `textplot_keyness()`

Keyness

```
key.analysis <- textstat_keyness(x = mig.dfm,target = 1)

textplot_keyness(key.analysis)
```


popul
shall
bnp
ethnic
percent
figur
minor
societi
multicultur
within
birth
statist
accord
social
group
third
on
offici
school
current
points-bas
fair
want
appli
detent
live
valu
forc
requir
migrat
need
can
visa
eu
end
student
control
new
migrant
system
-10
0
10
20
chi2
BNP
reference

Models

Correspondence analysis

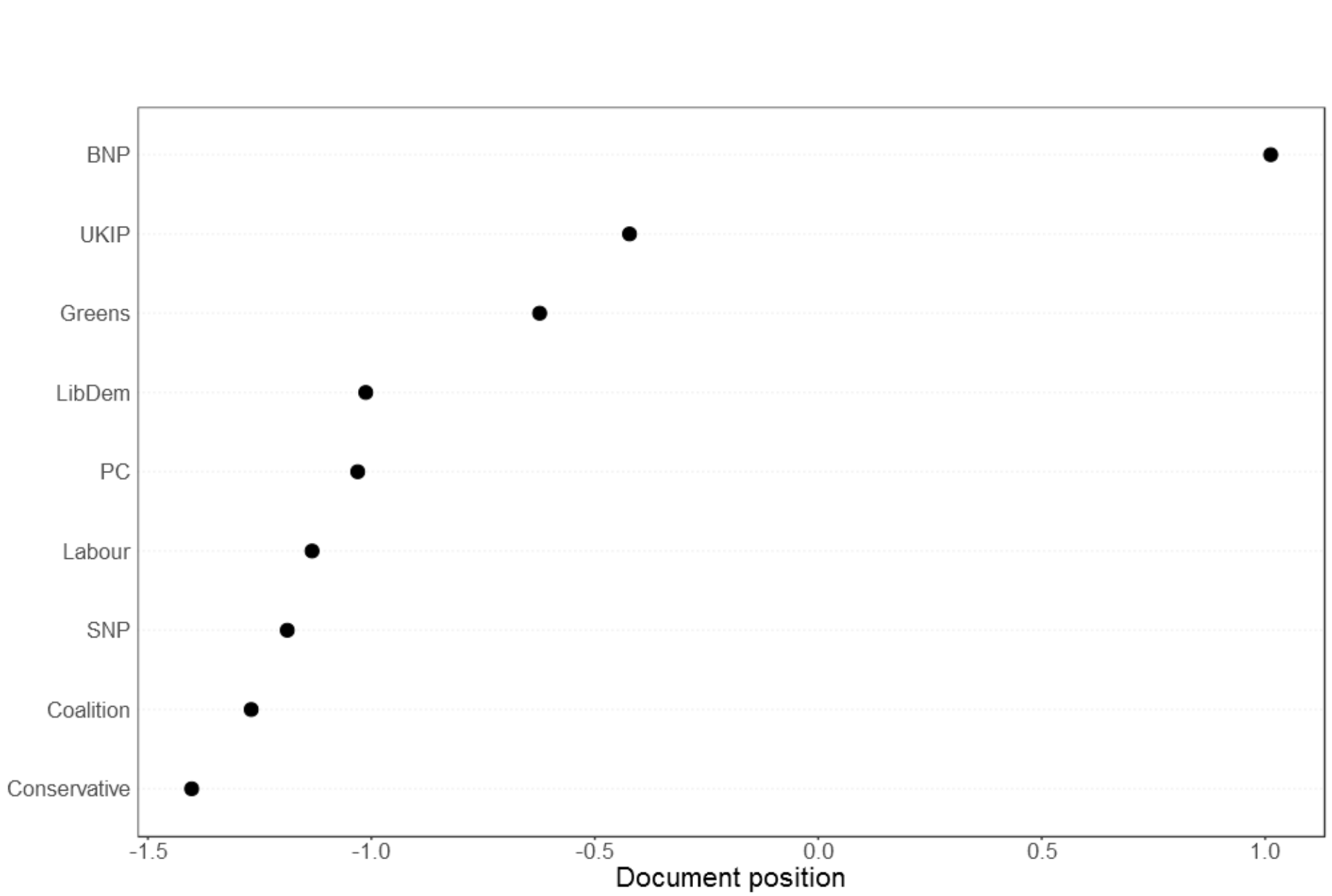
- Method of singular value decomposition
- Allows to reduce complexity of matrix into low-dimensional space (2 or 3)
- No underlying assumptions about distributions
- Scaling is a method of capturing the variation in the observed data
  - Not clear **what** is the variation captured (actual positions, tone, style, …)

Correspondence analysis

- Function `textmodel_ca()`
- Arguments
  - `sparse`
    - Allows to omit less frequent words in order to reduce the use of computer memory
  - `nd`
    - Default estimates as many dimensions as possible, allows to limit the number of estimated dimensions
- Useful to explore model with function `summary()`

Correspondence analysis

```
model <- textmodel_ca(mig.dfm,sparse = TRUE)

summary(model)

textplot_scale1d(model)
```


BNP
UKIP
Greens
LibDem
PC
Labour
SNP
Coalition
Conservative
-1.5
-1.0
-0.5
0.0
0.5
1.0
Document position

WordFish

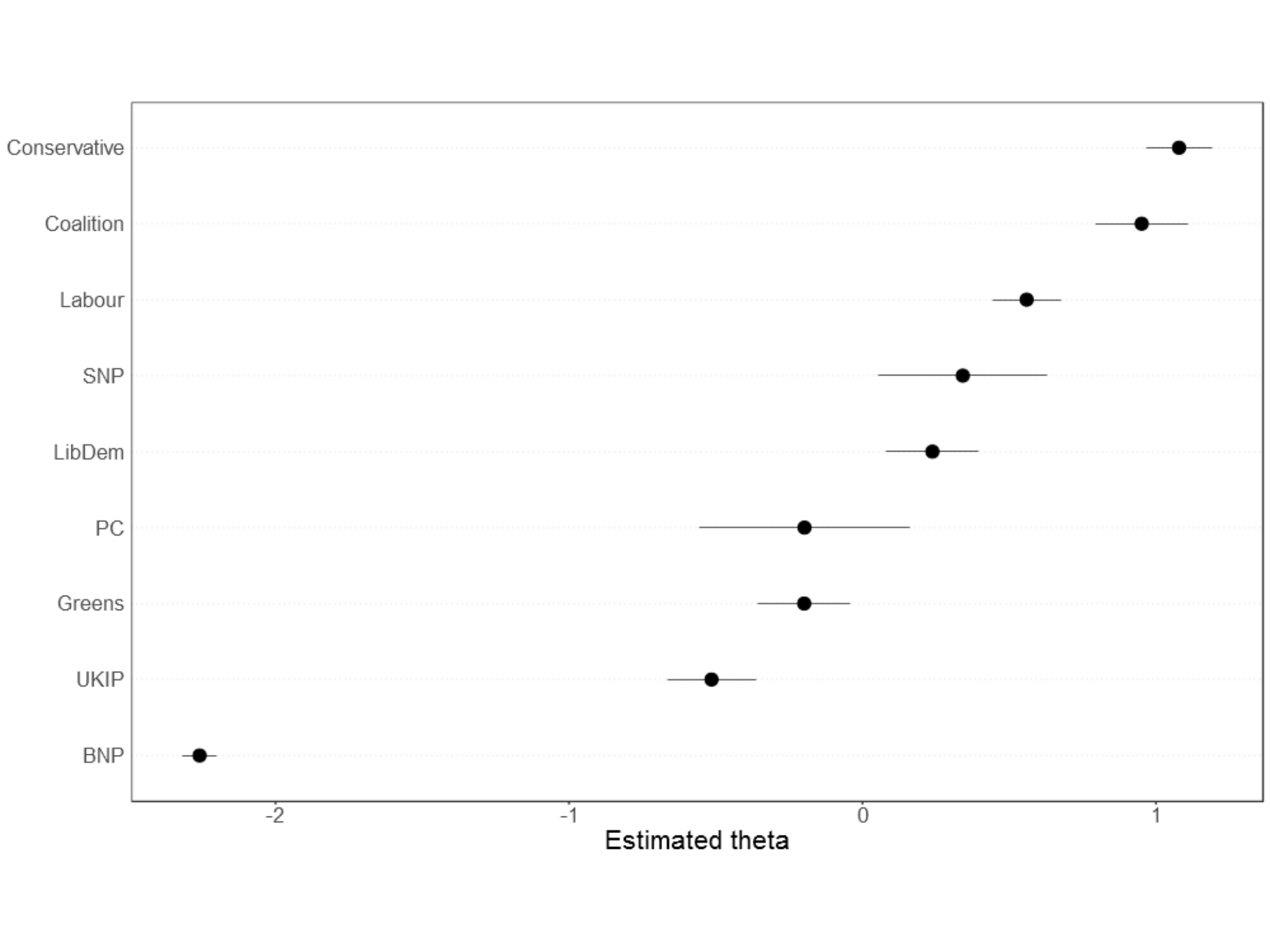
- Model based on naïve Bayes classifier
- Estimation of **one dominant dimension**
- Assumes a word is drawn from a Poisson distribution, which is based on
  - Amount the actor speaks
  - Frequency how much the word is used
  - Extent how much the word discriminates the underlying ideological space
  - Actors' underlying position
- Model is **estimated** given the observed data
- Again, lack of clarity, what the scale captures

WordFish

- Function `textmodel_wordfish()`
  - Arguments allow a further specification of prior assumptions about the Poisson distribution, model parameters, …
- Result provides also SE for each estimated position
- Function `summary()` allows to see the estimated model
- Function `textplot_scale1d()` allows to visualize results
  - Scaling of actors
  - Scaling of words using argument `margin`
  - Word highlight using argument `highlight` and a word list wrapped in function `c()`

WordFish

```
model <- textmodel_ca(mig.dfm,sparse = TRUE)

summary(model)

textplot_scale1d(model)
```

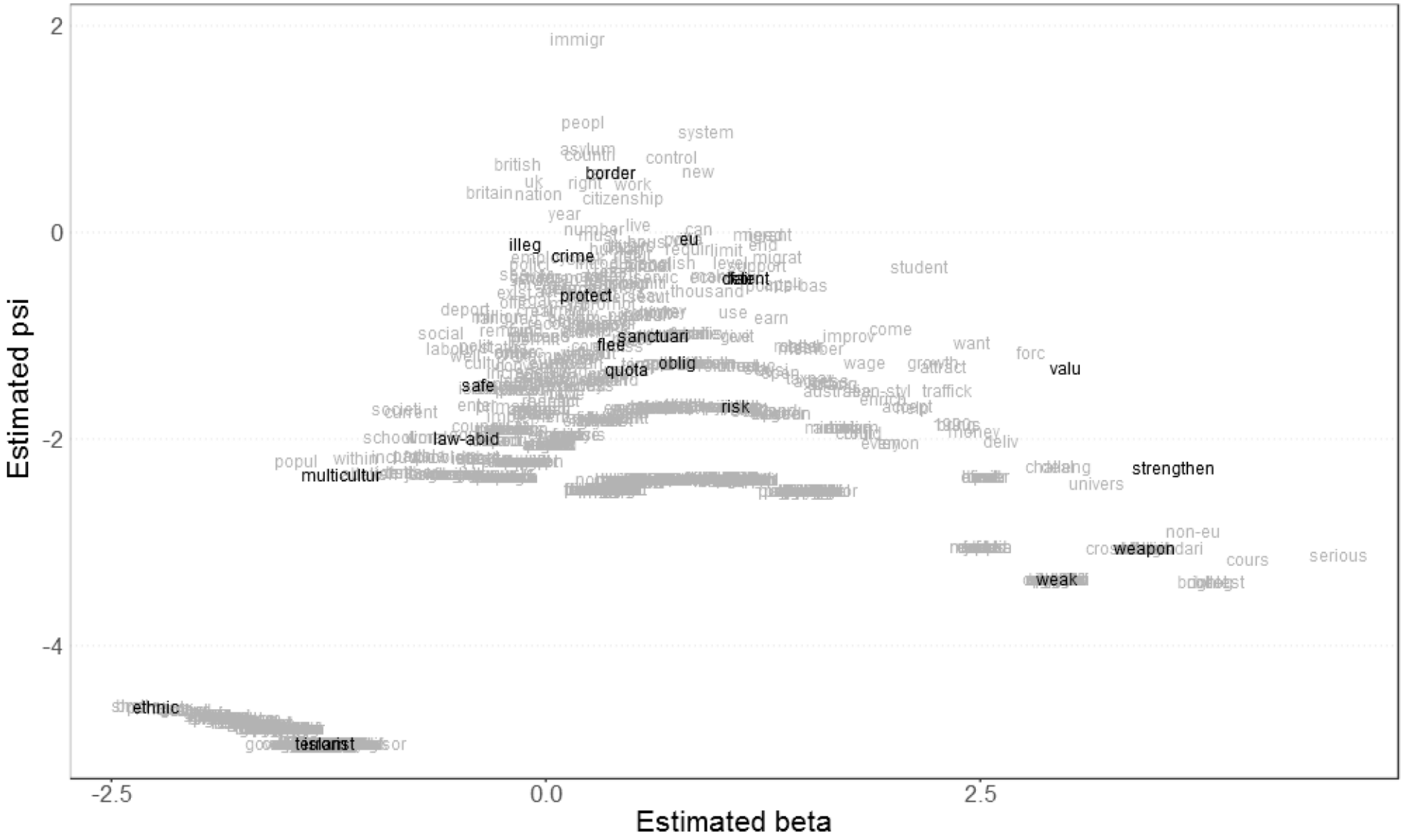

Conservative
Coalition
Labour
SNP
LibDem
PC
Greens
UKIP
BNP
-2
-1
0
1
Estimated theta

WordFish vs. CA

Wordfish	Correspondence Analysis
Conservative	Conservative
Coalition	Coalition
Labour	**SNP**
SNP	**Labour**
Liberal Democrats	**Plaid Cymru**
Plaid Cymru	**Liberal Democrats**
Green Party	Green Party
UKIP	UKIP
British National Party	British National Party

WordFish

```
textplot_scale1d(model, margin = "features")

textplot_scale1d(model,
                 margin = "features",
                 highlighted = c("eu","multicultur"),
                 highlighted_color = "black"
                 )
```


Estimated psi
Estimated beta
2
0
-2
-4
-2.5
0.0
2.5
immigr
peopl
system
asylum
countri
control
british
border
new
uk
right
work
britain
nation
citizenship
year
number
live
must
can
illeg
crime
eu
migrat
student
protect
thousand
use
earn
social
sanctuari
flee
quota
oblig
improv
come
want
wage
valu
safe
traffick
risk
law-abid
multicultur
strengthen
univers
non-eu
weapon
cours
serious
weak
ethnic
terrorist