Data Science Series Online Workshop

Topic Modeling for Text Analysis in R

Pre-Workshop Guide

November 2020

Contents

Introduction

This guide is a resource for students at Algoritma to use in setting up their laptop or environment prior to the scheduled workshops. In this guide, students can find a list of prerequisites that will be consistently used throughout the entire course. These prerequisites are required to be **completed before** the start of the workshop.

For new students, we will run through the installation process to ensure that the necessary programming languages and tools are installed. The next section will then talk about methods on how to verify whether the installs were completed successfully.

For recurring students, we recommend repeating the System Verification section once more to confirm past completed installations.

Installation

Installing R

Use this link: http://cran.us.r-project.org/

Installing RStudio IDE

Use this link: https://www.rstudio.com/products/rstudio/download/

Note:

1. Make sure the versions downloaded are compatible with your operating system.
2. Please make sure that you're installing R version 4.0.0 or above.

System Verification

Verify R Installation on Windows:

1. Type `R` in your Windows Search bar.
2. If the installation was completed successfully, there should be an R icon listed on the search result as shown below. In this case, it appears the user installed R version 4.0.0.

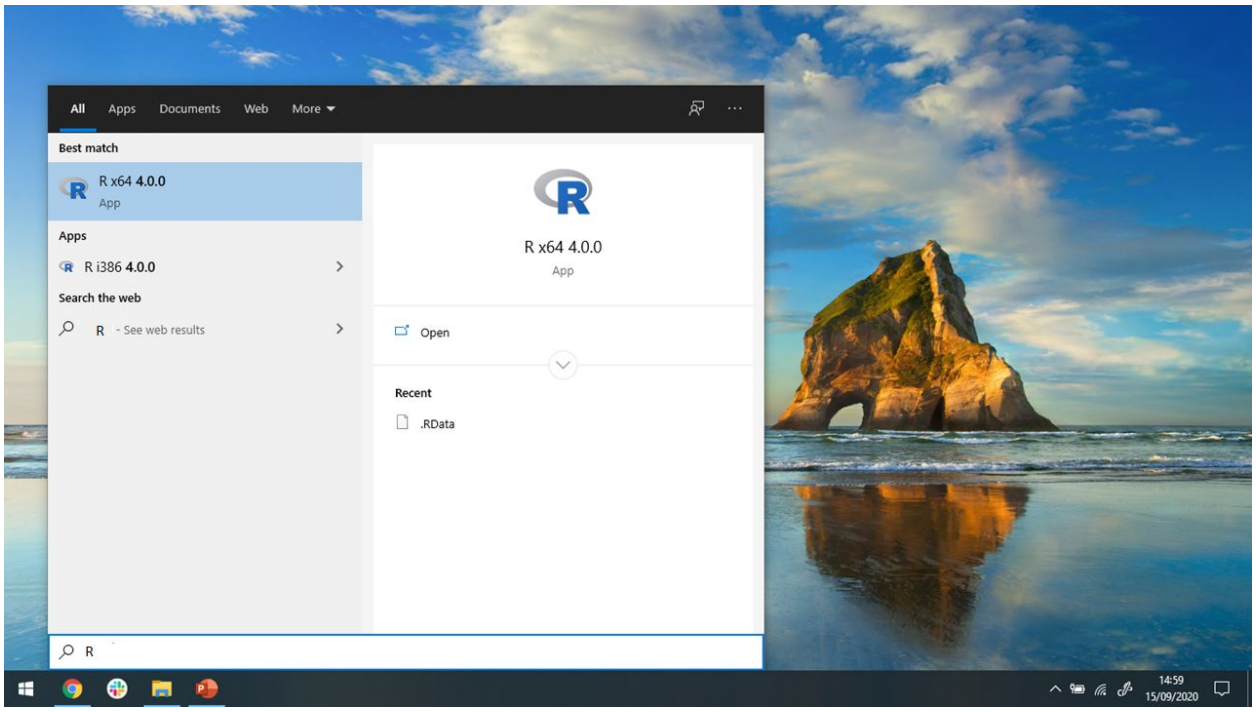


Figure 1: R icon on the Windows Search Bar result

Verify R Installation on Mac OS/Linux:

1. Type the command `R` in your console/terminal.
2. If the installation was completed successfully, there should be a response which includes information on which R version was installed as shown below. In this case, it appears the user installed R version 4.0.2.
3. To exit, enter the command q() and type:
    a. `y` - to save your workspace
    b. `n` - to NOT save your workspace
    c. `c` - to cancel termination of the process

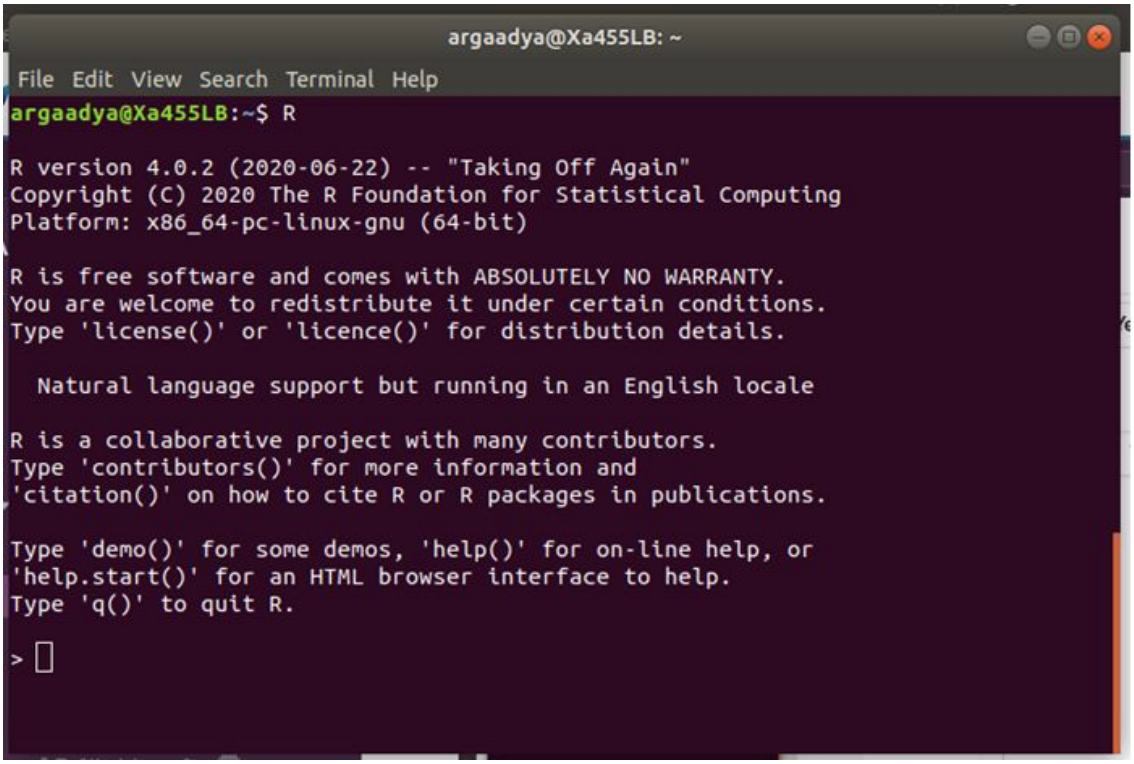


Figure 2: R response on Linux Terminal

Package Downloads

Downloading R packages can be done in the RStudio IDE in the "*Console*" box which is located on the bottom left-hand side of the screen. Simply type the command:

> install.packages("package_name")

whereas *package_name* refers to the name of the R package you would like to install.

Note: The quotation marks "" are part of the command so do not forget to include them.

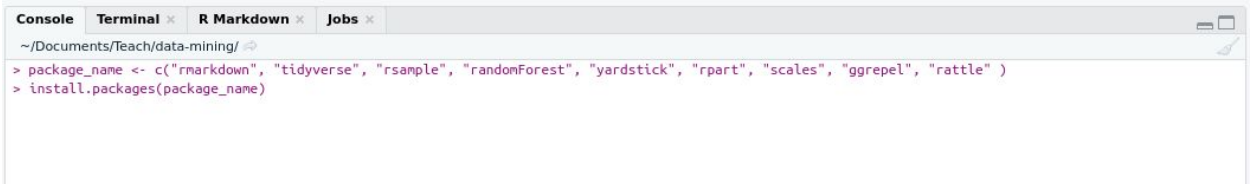
There are some packages that will be used in Solving Business Problem Using Data Mining Approach:

List the required packages

> packages <- c("rmarkdown", "ggplot2", "dplyr", "lubridate", "stringr", "tidyr", "tidytext", "SnowballC", "hunspell", "textmineR", "ggwordcloud", "scales")

Install the packages

> install.packages(package_name)
> install.packages("pacman")
> pacman::p_load_gh("trinker/lexicon","trinker/textclean")



During the installation, please make sure that you have internet connection and wait until the installation is complete as shown below.

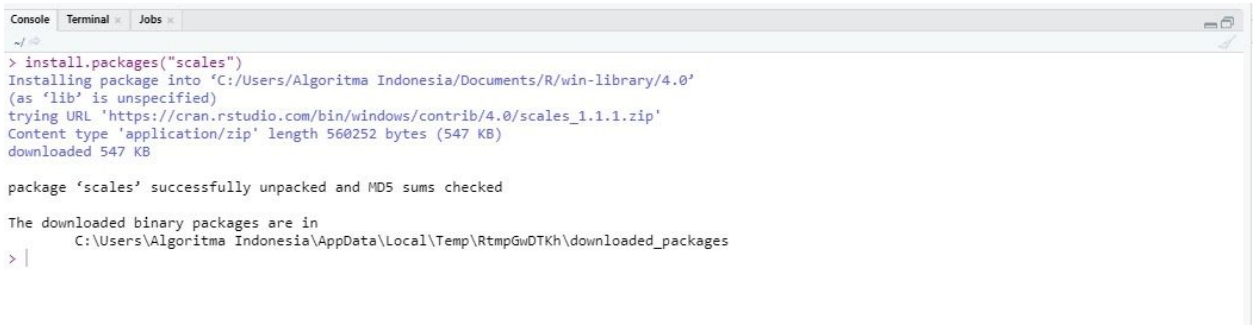


Figure 3: An example of a package installation

The installation process will take a lot of time, so please wait until the process is complete (after the '<' assign appears on the console).

To ensure that the packages already installed, you can move to the *packages* tab then type the name of each package on the search menu. If the package is available on the list so you have successfully installed the package. To verify your package installation and load the package into the R session, simply type the command:

```
> library(package_name)
```

Please run the following code to check if all packages are already installed. If there is no *error* message, you are safe to go and your package installation is complete.

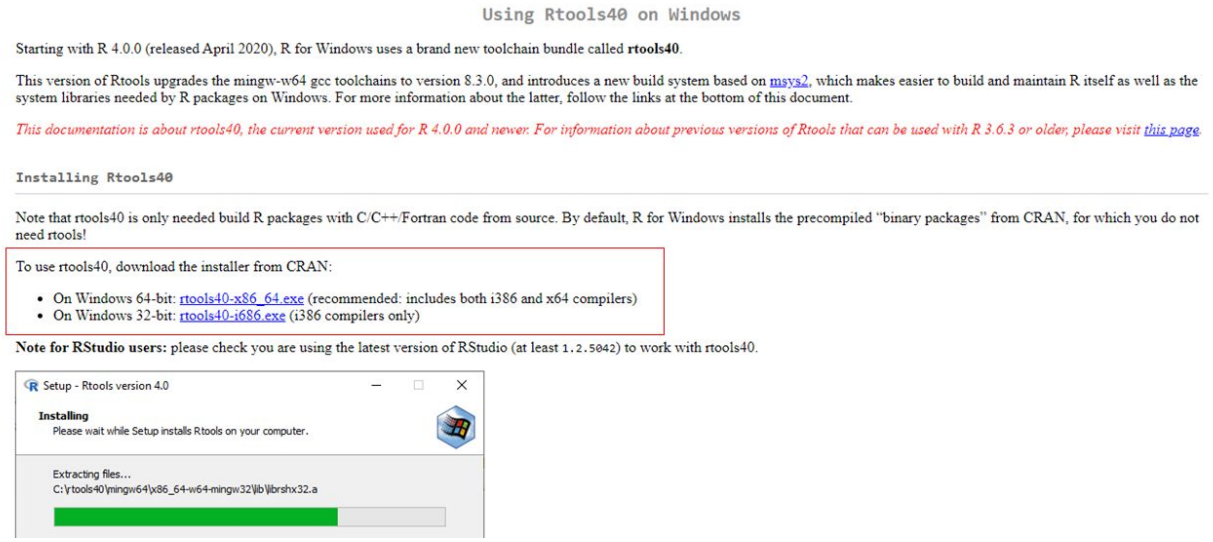
```
# Data Wrangling
library(dplyr)
library(lubridate)
library(stringr)
library(tidyr)

# Text Analysis
library(tidytext)
library(textclean)
library(SnowballC)
library(hunspell)

# Topic Modeling
library(textmineR)

# Data Visualization
library(ggplot2)
library(ggwordcloud)
library(scales)
```

Appendix

For Windows User - Rtools Installation

Some packages in R are written in C++ and cannot directly be used in R. You have to install Rtools to compile those packages.

1. Visit the following link: https://cran.r-project.org/bin/windows/Rtools/.
2. Download the installer based on your device system (64-bit or 32-bit).



Using Rtools40 on Windows

Starting with R 4.0.0 (released April 2020), R for Windows uses a brand new toolchain bundle called **rtools40**.

This version of Rtools upgrades the mingw-w64 gcc toolchains to version 8.3.0, and introduces a new build system based on msys2, which makes easier to build and maintain R itself as well as the system libraries needed by R packages on Windows. For more information about the latter, follow the links at the bottom of this document.

This documentation is about rtools40, the current version used for R 4.0.0 and newer. For information about previous versions of Rtools that can be used with R 3.6.3 or older, please visit this page.

Installing Rtools40

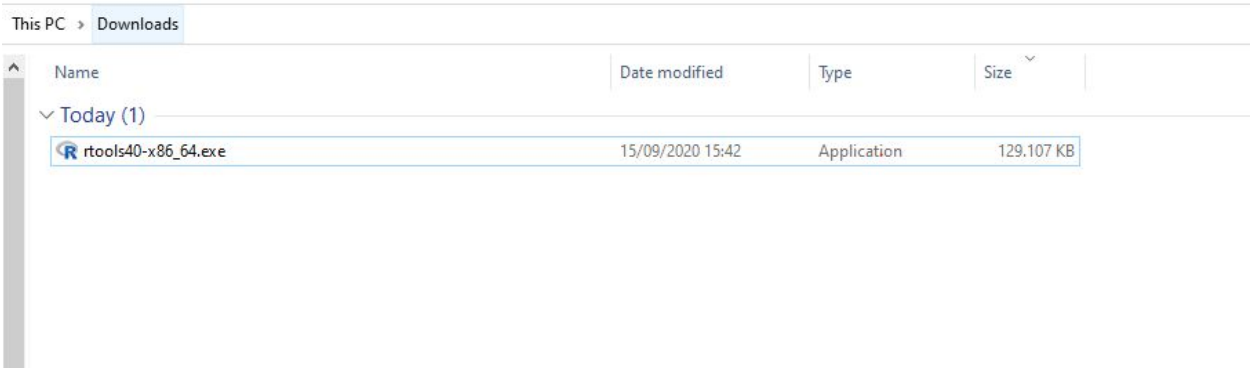
Note that rtools40 is only needed build R packages with C/C++/Fortran code from source. By default, R for Windows installs the precompiled "binary packages" from CRAN, for which you do not need rtools!

To use rtools40, download the installer from CRAN:

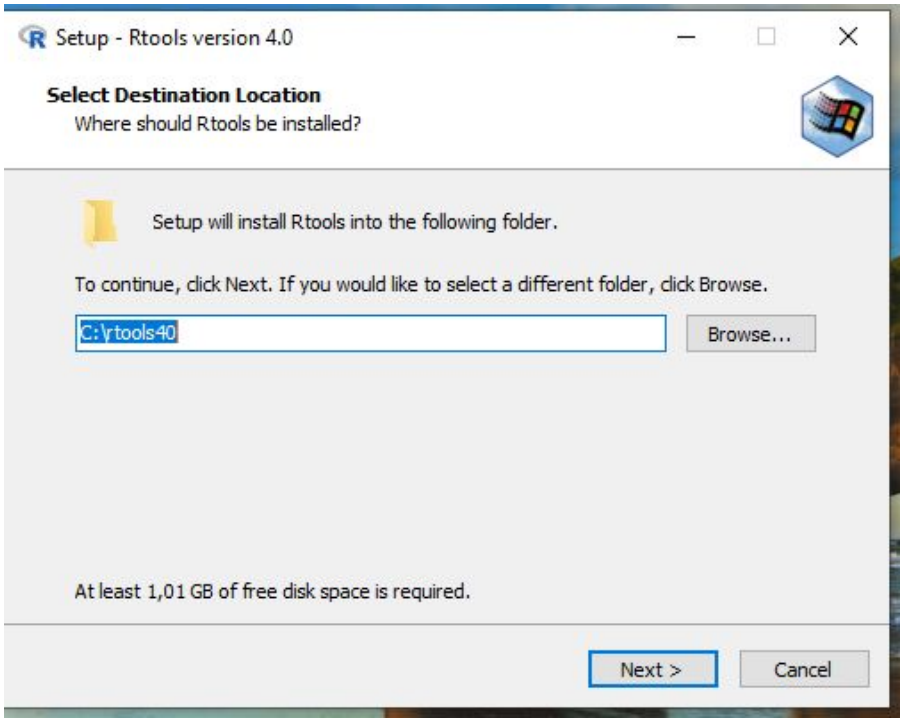
- On Windows 64-bit: rtools40-x86_64.exe (recommended: includes both i386 and x64 compilers)
- On Windows 32-bit: rtools40-i686.exe (i386 compilers only)

Note for RStudio users: please check you are using the latest version of RStudio (at least 1.2.5042) to work with rtools40.

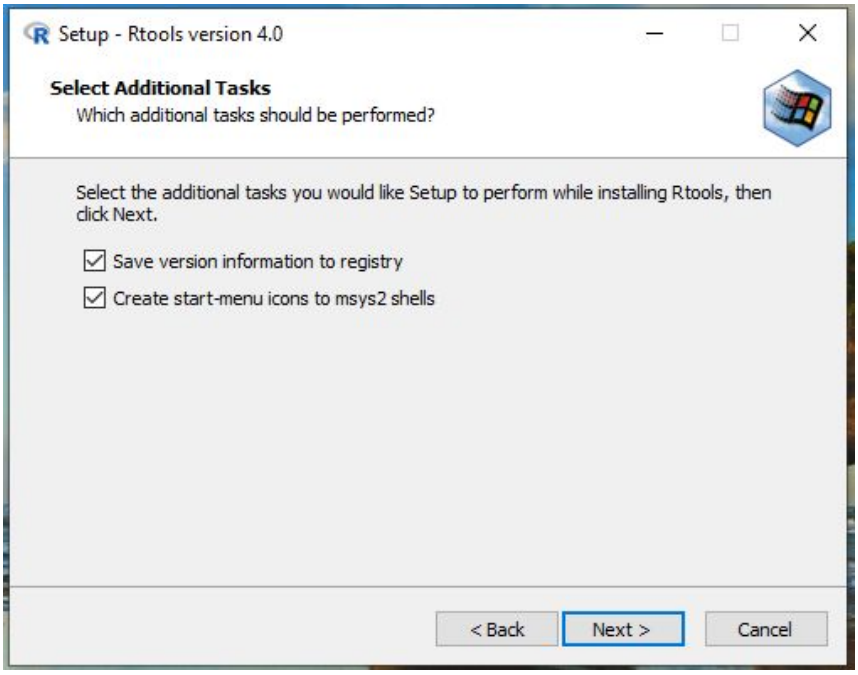
3. Run the installer *.exe* file.



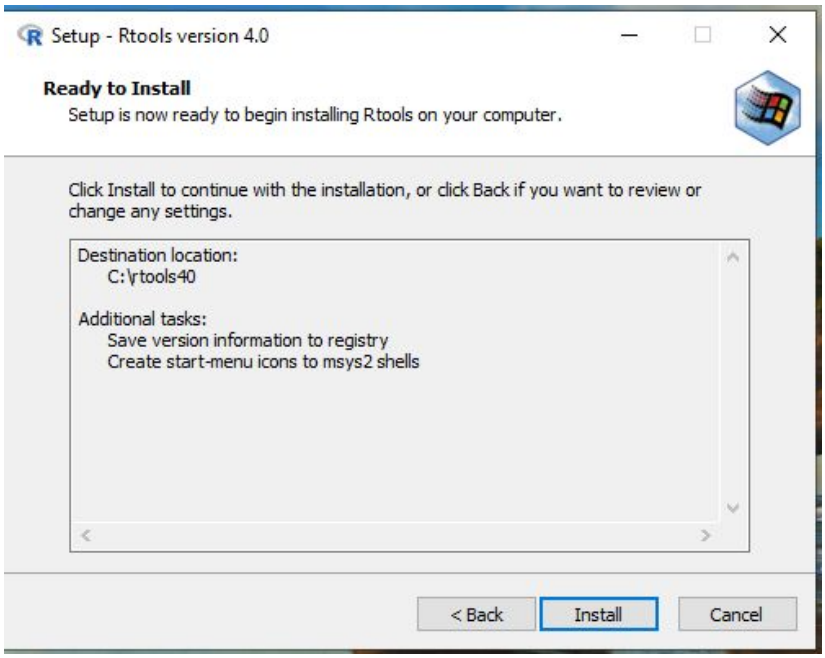
4. Click Next if the installation window has opened.



5. Click Next if you are already on this page.



6. Click Install to install the Rtools



7. Wait until the installation is completed and click Finish. Now you can use all available packages in R.