Dynamic Documents with R and knitr

Leonel Muñoz Cedano
leoneling@gmail.com

Universidad Distrital Francisco José de Caldas
Maestría en Ciencias de la información y las comunicaciones
Tendencias en Ingeniería de Software
Bogotá DC, 16 de Abril 2016

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Agenda

Introduction

- The basic idea behind dynamic documents stems from literate programming, a programming paradigm conceived by Donald Knuth (Knuth, 1984).

Introduction

- The basic idea behind dynamic documents stems from literate programming, a programming paradigm conceived by Donald Knuth (Knuth, 1984).
- Web scripting language such as PHP (which can embed program code into HTML documents).

Introduction

- The basic idea behind dynamic documents stems from literate programming, a programming paradigm conceived by Donald Knuth (Knuth, 1984).
- Web scripting language such as PHP (which can embed program code into HTML documents).
- Easy manage and maintain for the author.

Introduction

- The basic idea behind dynamic documents stems from literate programming, a programming paradigm conceived by Donald Knuth (Knuth, 1984).
- Web scripting language such as PHP (which can embed program code into HTML documents).
- Easy manage and maintain for the author.
- That is why we need the literate programming paradigm. In this paradigm, an author has two tasks.

Introduction

Technically, literate programming involves three steps:

- Parse the source document and separate code from narratives.
- Execute source code and return results.
- Mix results from the source code with the original narratives.

Introduction
About the Author

- Who is Yihui Xie.

Introduction
About the Author

- Who is Yihui Xie.
- Won the 2009 John M. Chambers Statistical Software Award (ASA).

Introduction
About the Author

- Who is Yihui Xie.
- Won the 2009 John M. Chambers Statistical Software Award (ASA).
- Honorable Mention prize in the Applications of R in Business Contest 2012.

Introduction
About the Author

- Who is Yihui Xie.
- Won the 2009 John M. Chambers Statistical Software Award (ASA).
- Honorable Mention prize in the Applications of R in Business Contest 2012.
- He founded the “Capital of Statistics”.

Installation

Installation requirements and configurations needid:

1. `install R x64 (version 3.2.3 or top)`
2. `install RStudio (version 0.96.331 or top)`
3. `install packages "knitr" in R commander`
4. `install TexStudio with MiKTeX`
5. `Integration RStudio with TexStudio`

Installation
Example Knitr - code

Example document *.Rnw in TexStudio

```
\documentclass{article}
\usepackage[utf8]{inputenc}
\begin{document}
```

Código de LaTeX normal, puede utilizar cualquier plantilla.

```
<<>>=
# Crea una secuencia de números
X = 2:50
# Muestra las medidas estadisticas basicas
summary(X)
@
\end{document}
```

Result of the example above

Código de LaTeX normal, puede utilizar cualquier plantilla.

```
# Crea una secuencia de números
X = 2:50
# Muestra las medidas estadisticas basicas
summary(X)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2      14      26      26      38      50
```

Reproducible Research

Characteristics document of Reproducible research:

- Results from scientific research have to be reproducible to be trustworthy.

Reproducible Research

Characteristics document of Reproducible research:

- Results from scientific research have to be reproducible to be trustworthy.
- Reproducible research (RR) is one possible by-product of dynamic documents.

Reproducible Research
Literature

- Proposed by Jon Claerbout at Stanford University (Fomel and Claerbout, 2009).

Reproducible Research
Literature

- Proposed by Jon Claerbout at Stanford University (Fomel and Claerbout, 2009).
- Sweave (Leisch, 2002) was among the first implementations for dealing with dynamic documents in R.

Reproducible Research
Literature

- Proposed by Jon Claerbout at Stanford University (Fomel and Claerbout, 2009).
- Sweave (Leisch, 2002) was among the first implementations for dealing with dynamic documents in R.
- The knitr package (Xie, 2015).

Reproducible Research
Good and Bad Practices

The key to keep in mind for RR is that other people should be able to reproduce our results. Some good practices for RR follow:

- Manage all source files under the same directory and use relative paths whenever possible.
- Do not change the working directory after the computing has started: function setwd().
- Compile the documents in a clean R session.
- Avoid the commands that require human interaction.
- Avoid environment variables for data analysis.
- Attach session Info and instructions on how to compile this document.

Reproducible Research
Barriers

Despite all the advantages of RR, there are some practical barriers, and here is a non-exhaustive list:

- The data can be huge.
- Confidentiality of data.
- Software version and configuration.
- Competition.

A First Look
Setup

- Since knitr is an R package, it can be installed from CRAN in the usual way in R:

```
install.packages("knitr", dependencies = TRUE)
```

- If you have any problems with knitr, the first thing to check is its version:

```
packageVersion("knitr")
# if not the latest version, run
update.packages()
```

- Once we have knitr installed

```
library(knitr)
knit("your-file.Rnw")
```

A First Look
Minimal Examples


```
\documentclass{article}
\usepackage[utf8]{inputenc}
\begin{document}
\title{A Minimal Example}
\author{Yihui Xie}
\maketitle
```

We examine the relationship between speed and stopping distance using a linear regression model: $Y = \beta_0 + \beta_1 x + \epsilon$.

```
«model, fig.width=4, fig.height=3,
fig.align='center'»=
par(mar = c(4, 4, 1, 1), mgp = c(2, 1, 0), cex = 0.8)
plot(cars, pch = 20, col = 'darkgray')
fit <- lm(dist   speed, data = cars)
abline(fit, lwd = 2)
@
```

The slope of a simple linear regression is

```
\Sexpr{coef(fit)[2]}
\end{document}
```

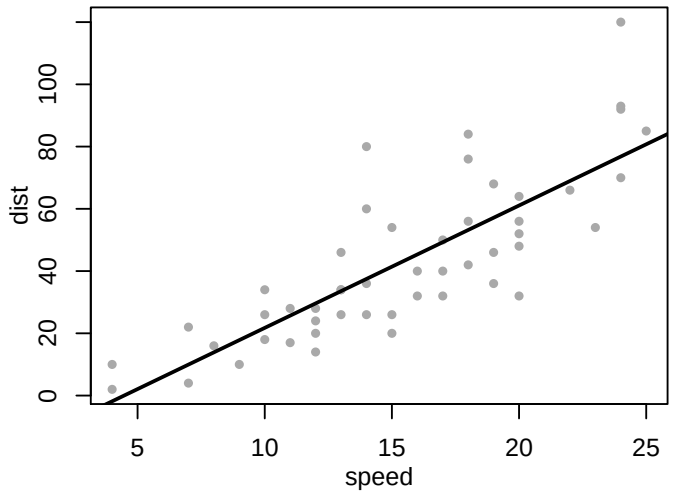
A Minimal Example

Yihui Xie

April 16, 2016

We examine the relationship between speed and stopping distance using a linear regression model: $Y = \beta_0 + \beta_1 x + \epsilon$.

```
par(mar = c(4, 4, 1, 1), mgp = c(2, 1, 0), cex = 0.8)
plot(cars, pch = 20, col = 'darkgray')
fit <- lm(dist ~ speed, data = cars)
abline(fit, lwd = 2)
```



The slope of a simple linear regression is 3.9324088.

If a user only has basic knowledge of R but knows nothing about knitr, or one does not want to write anything other than an R script, it is also possible to generate a quick report from this R script using the stitch() function.

The basic idea of stitch() is that knitr provides a template of the source document with some default settings, so that the user only needs to feed this template with an R script (as one code chunk); then knitr will compile the template to a report. Currently it has built-in templates for LATEX, HTML, and Markdown.

The usage is like this:

```
library(knitr)
stitch("your-script.R")
```

A First Look
Extracting R Code

For a literate programming document, we can either compile it to a report (run the code), or extract the program code in it. They are called “weaving” and “tangling,” respectively. Apparently the function knit() is for weaving, and the corresponding tangling function is purl() in knitr. For example,

```
library(knitr)
purl("your-file.Rnw")
purl("your-file.Rmd")
```

Feedback
Bugs, Comments and Suggestions

- Any Question, Comment or Suggestion.

Feedback
Contact Information

In case you have comments, suggestions or questions, please do not hesitate to contact me. You can find my contact details below.

Leonel Muñoz Cedano
leoneling@gmail.com

Bibliography

- Yihui Xie, Dynamic Documents with R and knitr, Second Edition, A Chapman and Hall Book, Taylor and Francis Group, 2015.
- Borbón A. and Mora W., Edición de Textos Científicos LaTEX 2014, Second Edition, Escuela de Matemática, Instituto Tecnológico de Costa Rica, 2014.
- `Pagina oficial R x64 (version 3.2.3) https://cran.r-project.org/bin/windows/base/.`
- Jose Nelson Pérez Castillo, Documento para Integración de RStudio con TexStudio, 2016.

Thank you!

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS