MPP-E1180 Lecture 5: Intro to Markup Lang. & Literate Programming (2)

Christopher Gandrud

25 September 2015

Objectives for the topic

- Review
- Advanced topics in markup languages and literate programming.
  - R Markdown Headers
  - Footnotes
  - BibTeX/Pandoc and citing R Packages
  - Time consuming analyses: caching and Make files

Assignment 2

Proposal for your Collaborative Research Project.

Deadline: Week 23

Submbit: A (max) 2,000 word proposal created with **R Markdown**. The proposal will:

- State your research question. And justify why it is interesting.
- Provide a basic literature review (properly cited with BibTeX).
- Identify data sources and appropriate research methodologies for answering your question.

As always, submit the entire GitHub repo.

Assignment 2

Definitely see me with your ideas/draft.

Start thinking about **types of models** that you want to use. I can include these in Lecture 8 (Statistical Modeling with R).

Review

- Why is literate programming useful for reproducible research?
- What is a markup language?
- What is a code chunk?
- What is the difference between Weave, Sweave, knitr, and R Markdown? (kind of a trick question)

Example Paper

An example of a paper + analysis + data project using many of the tools we cover today is available at:

HertieDataScience/Examples/SimplePaperWithAnalysis

R Markdown Headers

An R Markdown file is **just a text file** with markup instructions that **RStudio** understands.

The key to formatting is the **header**.

It is at the start of a file and comes between `---`.

The header is written in YAML.

YAML

YAML is a **human read-able data format**.

Elements are separated from values with a colon (:).

Each element is separated by new lines.

Hierarchy is maintained with tabs.

```
---
title: 'MPP-E1180 Lecture 5'
author: "Christopher Gandrud"
date: "9 October 2014"
output:
  ioslides_presentation:
    css: css/font-awesome.min.css
    logo: img/logo.png
  beamer_presentation: default
---
```

Super Nerd Point

YAML is a recursive acronym: ‘’YAML Ain’t Markup Language”.

Different Presentation Styles

By defualt, R Markdown uses the isoslides HTML presentation slides style.

You can also use reveal.js.

First install the revealjs R package:

```
devtools::install_github("jjallaire/revealjs")
```

Then in the YAML header use:

```
output: revealjs::revealjs_presentation
```

For further styling see
https://github.com/jjallaire/revealjs

Table of contents & Numbered Sections

You can add a table of contents and numbered sections to your PDF output:

```
output:
  pdf_document:
    toc: true
    number_sections: true
```

To do the same for **HTML** also include the information under `html_document`.

Figure Options

Create consistent figure formatting:

```
output:
  pdf_document:
    fig_width: 7
    fig_height: 6
    fig_caption: true
```

`fig_caption:` `true` attaches captions to figures.

To set the actual caption label, use the `fig.cap='SOME CAPTION'` **code chunk option**.

Pandoc footnotes

R Markdown can use Pandoc footnotes.

In-text: In the text place a **unique** footnote key in the format:

- `[^KEY]`

At the end of your document put the full footnote starting with the key, e.g.

- `[^KEY]: This is a footnote.`

BibTeX citations

BibTex allows you to create a **database** of **all** of the **literature/packages you cite**.

You can then insert them into your text and they will:

- Be **automatically formatted** consistently.
- Generate an appropriately ordered, consistently formatted **reference list** at the end of your document **with only the works you actually cited**.

The BibTeX Database

A BibTeX database is just a text file with the extension `.bib`.

Each entry follows a standard format depending on the type of media.

```
@DOCUMENT_TYPE{CITE_KEY,
    title = {TITLE},
    author = {AUTHOR},
    . . . = {. . .},
}
```

Note: Commas are very important!

The Cite Key

The cite key **links** a specific citation in your presentation document to a specific BibTeX database entry.

They must be **unique**.

It **does not matter** what order your BibTeX entries are in the `.bib` file.

BibTeX Articles

```
@article{Acemoglu2000,
    author = {Acemoglu, Daron and Robinson, James A.},
    title = {Why Did the West Extend the Franchise? Democra
            and Growth in Historical Perspective},
    journal = {The Quarterly Journal of Economics},
    year = {2000},
    volume = {115},
    number = {4},
    pages = {1167--1199},
}
```

BibTeX Books

```
@book{Cox1997,
    title={Making Votes Count: Strategic Coordination in th
           Electoral Systems},
    author={Gary W. Cox},
    year={1997},
    volume = {7},
    publisher={Cambridge University Press},
    address = {Cambridge}
}
```

More

For more media types and entry fields see
`http://en.wikipedia.org/wiki/BibTeX`.

Tip: Google Scholar

Google scholar generates BibTeX entries.

On an entry click `Cite > BibTeX`.

For a YouTube how-to see
`https://www.youtube.com/watch?v=SsJSR2b4_qc`.

Sometimes they need to be **cleaned** a little.

Linking your .bib file.

To link your `.bib` file to your RMarkdown document add to the header:

```
bibliography:
    - BIB_FILE_NAME.bib
    - ANOTHER_BIB_FILE_NAME.bib
```

Note: The files should be in the **same directory** as your R Markdown file.

Including BibTeX citations in RMarkdown

R Markdown uses Pandoc syntax to include a citation in-text.

General format: `@CITE_KEY`.

So if the cite key is `Box1973` then `@Box1973` will return Box and Tiao (1973) in the text of the presentation document.

Formatting In-Text Citations

Markup	Result
`[@Box1973]`	(Box and Tiao 1973)
`[see @Box1973]`	(see Box and Tiao 1973)
`[see @Box1973, 33-40]`	(see Box and Tiao 1973, 33–40)
`[@Box1973; @Acemoglu2000]`	(Box and Tiao 1973; Acemoglu and Ro
`@Box1973 [33-40]`	Box and Tiao (1973, 33–40)

Reference List

A reference list with the full bibliographic details of all cited documents will be **automatically created** at the end of your document.

Tip: Put a `# References` at the very end of your R Markdown document to have a section heading before the reference list.

Citing R and R Packages

Why cite?

- Give **credit** to the software authors (just like when citing literature).
- Enable **reproducible research**: identify which software you used and **which version**.

Citing R and R Packages

Base R way: print citation, copy BibTeX entry into your *.bib* file.

Cite R:

```
toBibtex(citation())
```

```
## @Manual{,
##   title = {R: A Language and Environment for Statistical
##   author = {{R Core Team}},
##   organization = {R Foundation for Statistical Computing
##   address = {Vienna, Austria},
##   year = {2015},
##   url = {https://www.R-project.org/},
## }
```

Citing R and R Packages

Cite R Packages:

```
toBibtex(citation('dplyr'))
```

```
## @Manual{,
##   title = {dplyr: A Grammar of Data Manipulation},
##   author = {Hadley Wickham and Romain Francois},
##   year = {2015},
##   note = {R package version 0.4.3},
##   url = {http://CRAN.R-project.org/package=dplyr},
## }
```

Citing R and R Packages: LoadandCite

The dynamic literate programming way: Use `LoadandCite` from the repmis package.

Load all of the packages at the beginning of you R Markdown file in a chunk with `include=FALSE`.

`LoadandCite` loads the packages and creates a BibTeX file with all of the citations.

```
pkgs <- c('dplyr', 'ggplot2')

repmis::LoadandCite(pkgs, file = 'RpackageCitations.bib')
```

Note: Use a file name that is different from your literature BibTeX file!

Citing R and R Packages: LoadandCite

Include the .bib file in your RMarkdown header.

Each **cite key** follows: `R-PKG_NAME`.

R itself has the key `CiteR`.

So `@R-dplyr` and `@CiteR` create the citations:

- Wickham and Francois (2015)
- R Core Team (2015)

Time Consuming/Intensive Analyses

Knitting your analysis and presentation documents together by placing all of your R code into code chunks can sometimes be **problematic**:

- When they are **time consuming** (requires a lot of computational time).
- When they **access files over the internet** (bad practice to make many repeated calls to the same URL, can crash the site. This is equivalent to a denial-of-service attack).
- When they are **many lines long**.

Solutions

- Long lines: use `source()` to run R code in other files.
- Caching for time/computationally intensive work: `cache=TRUE` code chunk option: only runs the chunk when the **chunk code changes**.

Make files

Make files are the ultimate solution to these problems.

Make is a command line program.

Big Idea: run a *make* file that runs a list of specific files in order.

Files are **only run** if they have been **changed** since the last time the make file was last run.

See Ch. 6 of RRRR if you might want to do this.

Seminar: Practice

Clone the HertieDataScience/Examples repo and play around with **SimplePaperWithAnalysis**.

Seminar: Begin working on your Proposal

Begin working with your partner on your research proposal.

- Identify the research area and key literature.
- Create a new repo and R Markdown document for your proposal.
- Begin building a BibTeX database for your key literature and try including them in your proposal.
- Begin identifying data sources.

References

Acemoglu, Daron, and James A. Robinson. 2000. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective." *The Quarterly Journal of Economics* 115 (4): 1167–99.

Box, G. E. P., and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Wickham, Hadley, and Romain Francois. 2015. *Dplyr: A Grammar of Data Manipulation*. `http://CRAN.R-project.org/package=dplyr`.