DSCI353-353m-453: Class 01a-p Open Data Science Tool Chain

2001-353-353m-453-00a-p-Open Data Science Tool Chain

Roger H. French, Peitian Wang

14 January, 2020

Contents

1.1.2.1 Literate Programing: Donald Knuth

Donald Knuth

- Bachelors and Masters degrees from CWRU
- PhD from CalTech
- CS Professor at Stanford

Did a great many things in Computer Science

- TAOCP: The Art of Computer Programming
  - Started in 1962, and not yet finished
  - Currently 7 volumes
- He also develeded TeX, the precursor to LaTeX

1.1.2.1.1 Literature Programming, was another of his goals

Literate programming is a programming paradigm introduced by Donald Knuth

- in which a program is given as an explanation of the program logic
  - in a natural language, such as English,
- interspersed with snippets of macros and traditional source code,
  - from which a compilable source code can be generated.

The literate programming paradigm, as conceived by Knuth,

- represents a move away from writing programs
  - in the manner and order imposed by the computer,

  - and instead enables programmers to develop programs in the order
  - demanded by the logic and flow of their thoughts.
- Literate programs are written as an uninterrupted exposition of logic
  - in an ordinary human language, much like the text of an essay,
  - in which macros are included to hide abstractions and traditional source code.
- Literate programming (LP) tools are used
  - to obtain two representations from a literate source file:
  - one suitable for further compilation or execution by a computer, the "tangled" code,
  - and another for viewing as formatted documentation, which is said to be "woven" from the literate source.
- While the first generation of literate programming tools
  - were computer language-specific,
  - the later ones are language-agnostic
  - and exist above the programming languages.

Now adays one can integrate R and Python code in a common shared environmet,

- as can be done with Rstudio v1.2 and the reticulate package.
- We use this in our data analytics in the SDLE Research Center at CWRU.

1.1.2.2 Your Open Data Science Tool Chain

1.1.2.2.1 Its all about a Data Science Tool Chain

- Use R and build on the communities foundation
- Use Rstudio as a comfy environment
- Share your Open Data and Open Source Code
- Produce Reproducible Science with Rmarkdown
  - Use Creative Commons Licenses
  - Or other Open Source Licenses
  - Such as the Gnu Public License: GPL

Pilot your DSCI studies using available data

- Find available data sets
- Before starting the costly process of making data

Use Git repositories

- For version control
- For Collaboration
- For Open Science sharing

1.1.2.2.2 Twitter used for Data Science

As part of setting up our Data Science Tool Chain

- Signup for a Twitter account
- Using Twitter in university research
- 10 Commandments of Twitter for Academics

Data Science People to follow on Twitter

- @hadleywickham
- @jtleek Jeff Leek JHU
- @rdpeng Roger Peng JHU

- @simplystats
- @Rbloggers
- @JennyBryan
- @hspter Hilary Parker
- @NSSDeviations
- @rstudio
- @rstudiotips
- @R_Programming
- @CRANberriesFeed
- @kaggle
- @SciPyTip
- @PyData
- @debian
- @ubuntu
- @GuardianData
- @UpshotNYT
- @EdwardTufte
- @ProjectJupyter
- @doctorow Cory Doctorow
- @gvanrossum founder of Python
- @NateSilver538
- @cutting founder of Hadoop
- @RProgLangRR
- @BitbucketStatus
- @CWRUITS_STATUS
- @cshirky Clay Shirky
-

1.1.2.2.3 Sign up for a Stack Exchange Account

Stack Exchange, Stack Overflow

- are a Q&A community focused on many topics.

Stack Overflow allows you to search by tag

- r and rmarkdown are useful tags for SO

Stack Exchange's Tour of Stack Overflow

Specific Stack Exchange websites

- for SX Data Science

- for SX Statistics on Cross Validated
- for SX Open Data

1.1.2.2.4 Efficiently browse you SX sites

- Google (but more random)
- The Stack Exchange apps
- Using an RSS Feed reader such as Feedly is a good way

1.1.2.2.5 Online Git Server Communities

- After your BitBucket Account
- You'll probably want a GitHub account,.
- Many Rprojects are there, and
- you can fork their repo's to inspect the code very easily.

1.1.2.2.6 Slack, another component of Agile Sofware Development

- cwru-dsci.slack.com
  - an online collaboration tool
- Its an intrinsic part of agile software development
  - There is slack app for phones
  - And client for computers, its on vdi.

1.1.2.3 You Online Data Science Portfolio

- Doing open, reproducible data science
- Lets you share a portfolio of codes and projects
- Cite it in your resume
- Build a community of supporters and collaborators
- Need to be conscious of data use terms and agreements
  - Funded research at CWRU falls under IP agreements
  - So when you consider licenses you want to use
  - They must be consistent with the IP terms that came
  - With datasets and codes

1.1.2.3.1 An Example, Emeline Liu

- emelineliu.com
  - This website, which runs off of Github Pages and Jekyll, is my latest project.
  - Right now, I'm using Poole as a foundation for my website/blog.

1.1.2.4 Links

- http://www.r-project.org
- Rory Winston, for the Learning R Intro
- StackExchange http://stackexchange.com/sites
- Twitter http://twitter.com
- HipChat http://hipchat.com
- emelineliu.com
- Github Pages
- Jekyll
- Poole