Measuring notebook reproducibility with repo2docker

Motivation

- A large-scale study of notebook reproducibility was published in 2019 [Felipe et al., 2019]
- repo2docker is a tool for automatically generating environments from repos, to aid in reproducibility [Project Jupyter et al., 2018]
- we want to evaluate repo2docker's automatic env generation, and compare with the study's conclusions
- Study did not evaluate execution of languages other than Python; repo2docker can produce environments with Julia, R, etc.
- repo2docker has no mechanism for evaluating success other than build completing

We built repo2docker-checker [**?**] to automate fetching and building repos with repo2docker, then running a test script to classify whether it should be considered reproducible.

Caveats

- For repo2docker, our main interest is in *environment reproducibility*, not *code reproducibility*, so if all dependencies are met, the goal is achieved. But how can we tell?
- **What qualifies as a successfully reproducible repo?** We chose: repo2docker successfully builds an image, and
  - it has notebooks
  - notebooks execute without error

Background

- reproduce previous study
- compare with repo2docker
- evaluate "automate existing best practices" claim
- failure to reproduce != repo2docker failure!

Process

1. collect repositories to test
   - source 1: github search for language:jupyter
   - source 2: queries from [Felipe et al., 2019]
2. build image with repo2docker
3. find notebooks in repo (up to 5)
4. execute notebooks with nbconvert

Categories of notebooks

Based on queries from previous study database, we collected and tested samples of repos, grouped by classification in the study:

- Julia notebooks (not tested previously)
- R notebooks (not tested previously)
- Installation failed (environment specified, could not install)
- Import error / ModuleNotFound (likely missing dependency specification)
- Success without dependencies (environment not found by study, execution succeeds)
- Success with default dependencies (study ran with anaconda, success)

Outcomes

Several possible classes of failure, some of which are interesting for repo2docker, others are not:

- build failure
  - due to partially pinned environment (e.g. pinned numpy not compatible with Python 3.8)
  - due to pre-build dependencies (e.g. apt-get install C libraries)
  - actually invalid requirements
- execution failure
  - environment not specified at all
  - environment insufficiently specified
  - user input required (e.g. credentials for using an API, or "fill out X before executing")

Results

Right	Left	Default	Center
12	12	12	12
123	123	123	123
1	1	1	1

Table: Demonstration of pipe table syntax.

Conclusions

- Open data publication in [Felipe et al., 2019] greatly facilitates follow-up studies!
- A large fraction of build failures were due to pinning packages (most often numpy or pandas) but not Python, because there is no community standard for pinning Python (conda, Pipfile insufficiently common, often lack Python itself)
- repo2docker success rate could be greatly improved if we looked at more of the repo to determine runtime language and/or version:
  - last repository commit date could pick older Python by default (not at all intrusive)
  - notebook `language info` metadata includes runtime and version (more expensive to evaluate)
- repo2docker does not implement best practices adopted in R and Julia communities
  - big caveat for Julia: most repos tested were older than the advent of the new `project.toml`

Future work

- Test repo2docker with different strategies for picking runtimes / versions
- Evaluate more repos, our sample size was small
- Plugin check to Binder for more thorough evaluation of success on build
- Engage with R / Julia community to investigate standards for specifying environments. **Is repo2docker missing something, or do these communities not have widely used environment standards?** Specifically, evaluate adoption of recently implemented `project.toml` specification.

Bibliography

João Felipe, Leonardo, Vanessa, and Juliana. Dataset of A Large-scale Study about Quality and Reproducibility of Jupyter Notebooks, March 2019. URL https://doi.org/10.5281/zenodo.2592524.

Project Jupyter, Matthias Bussonnier, Jessica Forde, Jeremy Freeman, Brian Granger, Tim Head, Chris Holdgraf, Kyle Kelley, Gladys Nalvarte, Andrew Osheroff, M Pacer, Yuvi Panda, Fernando Perez, Benjamin Ragan Kelley, and Carol Willing. Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. In Fatih Akici, David Lippa, Dillon Niederhut, and M Pacer, editors, *Proceedings of the 17th Python in Science Conference*, pages 113 – 120, 2018. doi: 10.25080/Majora-4af1f417-011.

Vilde Dille Øvreeide and Min Ragan-Kelley

Simula Research Laboratory

simula