BEST PRATICES

June 19, 2019

1 Diário de melhores páticas

Best Pratices Daily Adriano da Silva Unicamp - FEEC - Pós-graduação - RA 262097 Reprodutibilidade em pesquisa computacional - IA369Z - 1o. Semestre de 2019 Profa. Leticia Rittner

Este projeto foi concebido no intuito de atender aos requisitos do curso de reprodutibilidade em pesquisa computacional. Este projeto tem finalidade meramente educacional.

Este documento visa listar as melhores práticas percebidas por este autor/aluno ao longo do projeto para a criação de um artigo executável e reprodutível.

1.1 Visão sobre reprodutibilidade

Tendo em vista a grande quantidade de pesquisas e artigos que são produzidos, observando-se também a as dificuldades para a replicação das mesmas, a reprodutibilidade vem se tornando cada vez mais requisitada por publicadores e em todo o universo acadêmico. Uma pesquisa reprodutível é uma pesquisa cuja qual pode ser reproduzida por terceiros, em outro tempo e onde possa se obter resultados coerentes em relação aos obtidos na pesquisa original.

1.1.1 Preparação para o início do projeto

Domínio do tema Para iniciar qualquer projeto tenha bem definido, logo de início, os objetivos da pesquisa que será realizada. Parece óbvio, mas minha experiência indica a necessidade de frizar este ponto. Tenha ao menos uma idéia sobre os métodos que, ao menos de início, pretende empreender para a realização da pesquisa e os possíveis resultados esperados.

Fazer uma pesquisa reprodutível requer também fazer um artigo executável. Esta execução (do artigo) pode ser automatizada dentro do próprio texto do artigo (o que é, na medida do possível, desejável) ou pode ser realizada manualmente, conforme o artigo descreva o passo a passo. Após definido com clareza o tema e objetivos da pesquisa, procure conhecer o que é um artigo executável, saber para que serve e como manipular um. Se possível, tenha contato e tente manipular alguns. Tenha em mente que, quanto mais automatizado for a execução do artigo, melhor será para a reprodutibilidade dele.

1.1.2 Código

Linguagem de programação Considere sempre optar por ferramentas e linguagens de programação que já tenha o domínio. Lembre-se que aprender a fazer um artigo reprodutível já é um trabalho naturalmente complexo e requer bastante tempo e dedicação. O aprendizado de uma

nova linguagem de programação, concomitante ao aprendizado de outras ferramentas que certamente serão introduzidas, bem como dos próprios conceitos de reprodutibilidade, podem tormar um tempo maior do que o previsto. Não subestime.

- Considere também quem Python é uma linguagem bastante conhecida pela comunidade científica o que pode ajudar a conseguir alguma facilida de suporte.

https://www.python.org/

Literate computing A ferramenta Jupyter Notebook é excelente para utilização de pesquisas e acompanhamento das etapas passo a passo, principalmente em pequenos experimentos.

- A ferramenta é simples de usar, mas não subestime a complexidade do conjunto de funcionalidades que a ferramenta oferece e que certamente serão úteis para um bom artigo.

https://jupyter.org/

1.1.3 Documentação:

Dicas de como escrever o artigo

- Procure sempre escrever o artigo em um idioma que domine completamente, inclusive no caso da aplicação de termos técnicos que envolvam o projeto. Caso precise escrever em um idioma cujo qual não tenha o total domínio, tenha auxílio de alguém que o tenha e que ajude sempre a revisar todo o texto.
- Fique atento para a clareza do texto, tendo sempre em vista que um terceiro irá ler e tentar compreender, muitas vezes, um assunto complexo oriundo de um contexto que ele ainda possa não estar completamente inserido.

- Procure criar uma estrutura de diretórios organizada para armazenamento dos diversos tipos de arquivos que compõe a pesquisa. Crie pastas com nomes simples e que aludam ao significado dos arquivos que estão dentro. Exemplos:
  - Pasta "data" para armazenamento dos arquivos de dados que serão utilizados na pesquisa.
  - Pasta "dev" para armazenamento dos arquivos de código fonte que manipulam e preparam dados da pesquisa.
  - Pasta "deliver" para armazenamento dos arquivos de código fonte que relatam ou entregam os dados em forma de artigo, bem como os arquivos de entrega do artigo propriamente ditos.
  - Pasta "daily" para armazenamento de um arquivo de diário de boas práticas.
  - Pasta "image" para armazenamento de imagens utilizadas no projeto.
  - Pasta "env" para armazenamento de arquivos de configuração do ambiente.
  - Pasta "var" para armazenamento de variáveis serializadas no projeto.

- Não esquecer de anotar o link dos sites onde são encontradas dicas para posteriormente colocar nas referências.

Link para o leitor do artigo iniciar a reprodução do mesmo Não esqueça de colocar no texto final do artigo, a indicação de um link ou descrição de onde o leitor poderá encontrar os arquivos para que a experiência possa ser baixada e/ou reproduzida (um link para uma página específica de um site onde esteja hospedada a experiência, um repositório para download). Lembre-se que é à partir do texto final do artigo que o leitor iniciará a busca por informações para a reprodução da experiência.

Ferramentas para documentação Aqui também é importante destacar a ferramenta Jupyter Notebook que pode, além de auxiliar na execução das pequenas tarefas do experimento, pode ajudar na documentação. Abre-se um parêntese aqui para destacar plugins como o nbconvert, pdflatex e bibitex que ajudam a automatizar a geração do artigo em PDF.

Dicas: - dedique um tempo para pesquisar essas ferramentas e seus funcionamentos; - Conhecimento sobre o editor latex pode ajudar;

http://theoval.cmp.uea.ac.uk/~nlct/latex/pdfdoc/pdfdoc/pdfdoc.html

Arquivo README Crie um arquivo README e coloque as instruções em um passo-a-passo minuncioso para que o leitor possa realizar a reprodução da experiência. Indique o caminho completo e por qual o arquivo ele deverá iniciar a leitura e em que ordem e quais arquivos deverá abrir e ler mais instruções até que consiga iniciar e realizar toda a experiência. Indique também dentro deste arquivo README, informações sobre os arquivos de licença de software e de dados do projeto, bem como qualquer outra informação adicional relevante para o leitor.

1.1.4 Testes

Após o término do artigo, faça testes de reprodutibilidade, tentando reproduzir o artigo passo a passo conforme descrito nos textos e tutoriais que tenham as instruções que auxiliam a reprodução. Faça esses testes em computadores diferentes do utilizado para o desenvolvimento das pesquisas. Neste tipo de teste é possível perceber pontos de falhas na reprodutibilidade.

1.1.5 Ambiente e Distribuição

Docker O Docker é uma ferramenta baseada em container de sistema operacional e pode ser muito interessante para a criação de um ambiente para execução de um experimento em diferentes computadores, o que o torna uma ferramenta muito útil para a reprodutibilidade.

Com o Docker você empacota todas as ferramentas e bibliotecas utilizadas e até mesmo os dados podem ser armazenados no interior de um contaiener, facilitando a vida de quem irá reproduzir o experimento.

O Docker possui ainda a ferramenta Docker Compose, que facilita ainda mais o gerenciamento de vários containers que podem ser necessários ao experimento e simplifica as execuções.

Links: https://docs.docker.com/install/ https://docs.docker.com/docker-for-windows/install/ https://docs.docker.com/docker-for-mac/install/

- Docker Compose: Em relação ao Docker Compose, tenha cuidado pois a versão do arquivo YML de configuração depende da versão do seu docker-engine, sendo que muitos comandos só existem à partir de determiada versão do arquivo YML, não estando disponíveis para todos. Vale ressaltar também atenção para caminhos relativos. O comando "-security-opt seccomp:./chrome.json", por exemplo, não é aceito pela ferramenta por conta do caminho relativo, obrigando a colocar um caminho absoluto para a localização do arquivo json, o

que amarra o ambiente e dificulta a distribuição. Além disso, atente para a necessidade de instalar a ferramenta docker-compose no ambiente onde for rodar o container através da ferramenta. No caso desta pesquisa, foi mais interessante rodar o ambiente através do comando docker run e não pelo Docker compose.

Git Hub O repositório do Git-hub e a ferramenta git auxiliam no controle de versão do software, além de possbilitar a distribuição de todo o experimento, inclusive os dados.

Dica: - Fique atento para o tamanho dos dados que serão armazenados em repositório na internet;

https://github.com/github

1.1.6 Dados

Deve-se buscar ao máximo alternativas para se conseguir que os dados das pesquisas estejam disponíveis para quem precisar reproduzir. Isso pode significar que os dados devam estar armazenados em algum repositório da internet.

Conforme já dito em outros tópicos, os dados podem também estar armazenados em containers Docker (e assim ficar no próprio Docker-hub, quanto também diretamente no Git Hub. Vale ressaltar apenas a necessidade de se observar os limites de armazenamento destes repositórios, bem como a dificuldade de trafegar grandes quantidades de dados pela internet.

1.1.7 Workflow

O Workflow é um recurso importante quando se está a realizar pesquisas relativamente grandes, de onde se necessita executar vários experimentos em diferentes etapas e tais ferramentas podem ajudar na visualização e até no orquestramento da execução de cada passo.

Entretanto, em pesquisas menores e com poucas etapas, pode ser um recurso dispensável.

Uma ferramenta válida para o desenho de workflows é o Draw.io.

https://www.draw.io/

```
In [ ]:
```