Data Science for AAE

AAE 718 – Summer 2024

Worksheet 6

GitHub

1 Reading

This worksheet, I guess. I'm also providing a large number of links. In general it can be surprisingly difficult to find a GitHub tutorial for somebody that knows nothing.

2 Daily Goals

- Understand how to use Git and GitHub
- Understand pushing, pulling, branching, pull requests and forking
- Merge conflicts

3 What is Git and why should you care?

Imagine you are working on a large coding project. This project has several (potentially hundreds) of files with thousands of lines of code. Moreover, people are using this code in their work. If you start making changes it's possible you'll break something vitally important and then people will be mad at you.[1]

This is where Git comes into play. You can tell Git to *branch* your code. This will leave your old code unaffected and you can make your changes in a safe environment. When you're confident everything works as expected Git will allow you *merge* your changes.

3.1 Git vs GitHub

By default Git is a local program, this means it lives on only your machine. But that's not how we share code. Hence GitHub, which is a cloud based Git implementation with many nice features. GitHub is *not* the only cloud based Git service, there are many. However, GitHub is widely regarded as the best. GitHub is also now owned by Microsoft, which has certain advantages when working with the US government[2].

4 Getting set up

Really all you need is GitHub Desktop. Here is a link to download what you need, you can also just Google it's all the same.

You'll also need a GitHub account. Click this link and get started. I would highly recommend making this using a personal account. You can always add other email addresses. As an example, I have been using GitHub for many years and I've lost email addresses from old institutions, if I had used one of those it would be quite annoying to recover my GitHub.

GitHub will probably also prompt you to set up the mobile app. This is a good idea for the dual authentication. GitHub security is *very* important. If your account gets compromised people can add malicious code into your projects. This is bad.

5 An Overview

First, this may seem relatively abstract at the moment as the practical examples will be below. But I want you to get an idea of what's happening. You can always reread this later. I'll also put this into a paper format with useful definitions highlighted.

[1] This should be avoided at all costs.

[2] Microsoft is the official cloud provider for the US government.

Definition 5.1 (Repository)**.** A *repository* is your code. When something references a *repository* they are talking about your GitHub, but you can also have local repositories. These are sometimes referred to as a *repo* for short.

I have a large number of repositories in GitHub, here is an example. This is the main repository for WiNDC, the thing I do for a living. Repositories can be either private or public and you can add people repositories so they can make changes.

Definition 5.2 (Clone)**.** When you *clone* a repository, you download it to your computer.

Cloning a repository is the first step to making changes to a repository *that you own*. You can clone any public repository, but you may not be able to change that code on GitHub.

Definition 5.3 (Fork)**.** A *fork* is a copy of somebody elses repository in your personal GitHub.

If you fork a repository, that is a copy that you own. You can clone that repository and make changes.

Definition 5.4 (Commit)**.** A *commit* is registered change to the code. You are required to include a small comment when you commit.

At it's core, Git is a sequence of commits. The power of Git is that you can easily trace back and undo a sequence of commits. In general, it's better to have many small commits vs one huge commit.

A commit is a local change only. If you commit some code on your machine, nobody will see it on GitHub.

Definition 5.5 (Push)**.** A *push* uploads your commits to GitHub.

Definition 5.6 (Fetch)**.** A *fetch* downloads commits from GitHub.

Definition 5.7 (Pull)**.** A *pull* fetches changes and makes them available to you.

In theory there is a difference between fetch and pull, but in practice it acts the same. Plus, whenever you push to a repository you'll also be pulling. In GitHub Desktop it's a single button.

Definition 5.8 (Pull Request)**.** A *Pull Request* takes all of the commits you've made and requests they be pushed into a "different" repository.

Pull requests are a GitHub feature, and they are incredibly useful. First, different is in quotes in the definition because I routinely pull request from my personal repository into itself. We'll discuss why this is so useful later.

When you open a pull request, you're requesting that the owner perform a pull from your repository. This can be dangerous for the owner, what if you're inserting malicious code? To counteract this GitHub has a built in code-review feature. You can inspect line-by-line every proposed change, you can even leave comments on individual lines of code. The owner of the repository may request you make changes, or might make the changes themselves.

Pull requests are how people contribute to open source software.

Definition 5.9 (Merge)**.** A *merge* is anytime code from one repository is being combine with another repository.

In the best case scenario, you won't have an issue merging. However, it will happen that code from one repositories conflicts with code in another, this is a merge conflict and they can be difficult. There are tools to help mitigate this, we will only scratch the surface.

Definition 5.10 (Branch)**.** A *branch* is a parallel version of your code. As you switch branches your code will change to reflect the new branch.

Branches are incredibly useful. If you have a project and you're working on several new features, each feature should live in a branch. This means you can focus on just your change and won't get distracted if another feature breaks something.

When you are finished with your branch, you open a pull request to your main branch so you can merge the changes. You can then delete your branch. The pull request process will generate automated patch notes based on your pull request. I can't overstate how invaluable this is. Here is an example, if you scroll down to "Merged Pull Requests", you'll see what I'm talking about.

6 How I typically use Git/GitHub

If I'm working on a project that's just me or has just started, I push commits directly to my main branch. A small code base makes this easy.

Once my project starts to get large, I'll make branches. In general, you should only branch from your main branch. Otherwise you can get issues where things get out of sync. Of course, this is not a rule and people do this all the time. But when you're first learning, it's best to keep things simple.

If I have an idea and want to test it, I'll make a local branch and just play with my idea. If it works out, I'll push the branch to GitHub and start the PR[3] process. If not, I can just switch back to my main branch and everything is back to normal.

7 Practical Tutorials

This may be seen as a cop-out on my part, but there are WAY better tutorials than I could every write.

- First contributions in GitHub Desktop
- Same as the above, but using tools directly in VSCode. This is what I do, but isn't necessary.
- Lots of tutorials. This is a lot of tutorials. You should do at least:
  - Introduction
  - Markdown (if you don't know it)
  - Review Pull Requests
  - Merge Conflicts

8 .gitignore

The last important thing to know: Git only likes raw text files. Uploading binary files (like PDFs or some image files) is discouraged. However, sometimes your code will automatically generate these types of files which means Git will see them. The solution is *.gitignore*.

In your local git repo there will be a file called *.gitignore*. Anything in this file will be automatically ignored by git[4]. You can put any regular expression in this file[5]. If you want to ignore all PDFs just put

```
*.pdf
```

in gitignore and save it. You may need to commit only the gitignore file, but then you should see the PDFs disappear from the git commit screen. Unless you've already uploaded a specific PDF, then it'll still be captured. There is another process to remove it which we won't cover here.

When you create a new empty repository there is typically a gitignore option that will create a default file for whatever language you're using.

[3] Pull Request
[4] Hence the name.
[5] We'll cover these later.

9 README

Every Git repository should have a README.md file, in fact you can should one in every subdirectory too. On GitHub if you scroll down on any repository you'll see text describing the repository, this text is in the README file. The README file is *markdown*, we'll discuss markdown in a future worksheet, but for now think of it as raw text, you can just type and it'll show up.

10 Summary

These tools should get you started on GitHub. You may not find this useful at the moment, but someday you might. Git can be useful anytime you're working with and share raw text files (like code as opposed to a PDF). Being able to sync changes with other people is huge.

GitHub is *way* deeper than we are going in this course. I have a Julia project set up in GitHub that whenever I push a change, GitHub will run tests in several operating systems and different versions of Julia. It will also build my documentation and push it to a website.

If you want to learn more about Git/GitHub, Google is a great tool. I've written this worksheet because most of the tutorials I've seen don't explicitly define these things, which is annoying. You should now be set to do most of what you need to do.

Writing homework for this is quite difficult. Most of these problems can't reasonable be graded by myself. We're going to have to go with the honor system for most of these.

For Problems 1 – 4 you'll be creating a repository and using it throughout.

Problem 1 (10 pt) Create a GitHub account (if you don't already have one) and download GitHub desktop. If you're feeling adventurous you can also get Git working directly in VSCode[6].

1. Create a new empty repository on your PC
2. Add some files to it, code from a previous worksheet would be fine
3. Commit these files and push the repository to Github
4. Go on Github and see your files

Problem 2 (10 pt)

1. Create a branch of your repository, you can do this directly in Github desktop[7]
2. Make some changes to your code and commit
3. Publish your branch to Github, basically do a push
4. You can now see your branch on Github
5. In your local repository, change back to your main branch and read through your code. You should see the changes you've made have disappeared.

This is the purpose of a branch. Make changes without destroying working code.

Here is a use-case example. Let's say I gave a problem where you get 5 points for having a working solution, but 10 points for the fastest solution. You should try to solve it first and get the 5 points, but then you want to try for the 10. You don't want to lose the code for the 5 points, so you make a branch.[8]

[6] This is how I use Git

[7] Probably covered in the tutorials

[8] Don't say "I'll just copy/paste my old code somewhere else". You will go confused eventually, especially if you're working in a team or have 5 copies. I'm speaking from experience working with someone who will make a small change and email it to you and then you have 7 copies of code with slight differences and you don't know which one is correct. Then you get to combine these into a single thing, but there is NO documentation on the changes or what has changed. So you do this and you miss things, because how would you know all the differences? Then your whole work flow slows down because the code no longer works and you can't remember what the original file was, was it "code-new", "code-original", "code-new2"? You

Problem 3 (10 pt)

1. Create a pull request to update your main branch with your new branch. You do this on GitHub.
2. Go to the Pull Requests tab and examine your PR. You should see four tabs, Conversation, Commits, Checks, Files Changed.
3. Conversation - You'll see your commits and comments. Make a new comment.
4. Commits - A better view on the commits
5. Checks - You won't see anything because we aren't covering checks. In my work I have GitHub configured to automatically run a number of tests. This screen will tell me if they have passed or failed.
6. Files Changed - You can actually see the changes and make comments on specific lines of code. You can mark a file as "Viewed" to collapse it. This is the first step of a quality Code Review. This is an information dense screen, but it fairly self-explanatory you should familiarize yourself with it.
7. Leave a comment on a few lines of your changed code. If you hover over the code you'll see a blue plus, you can click and drag to select multiple lines.

Problem 4 (10 pt)

1. Open an issue in your GitHub repo (in the issues tab)
2. Call it whatever you want and write a description. A good description in an actual repository[9] will have a brief description of the problem, a minimal working example, and, if you can, a potential solution. Here is an issue I wrote for Julia JuMP, you can see this is an open issue with lots of discussion.
3. Link your issue to your pull request. This is easy to do, in your Pull Request you can either edit the description or add a new comment and type a # and it should pop-up a list of all the issues. If you know the number you can type it, you can see the number on the issue page.

Issues are great. If they are linked to a pull request, when you merge the PR you can automatically close the issue. Plus if you release your code GitHub will list the new PRs and the issues they close, with links to each.

Problem 5 (10 pt) Here is a repository I created link to the repo. You have view access to this, because it's a public repository, but you can not edit it.

1. Fork this repository to your account
2. Add your name to `roster_names.md`, follow the given format
3. Open an issue on my repository with some name and description
4. Open a PR using your fork and link your issue in the description.

have no idea. Just force this person to use GitHub and make branches.

[9] In other words, you don't need to do this here