Python Crawling

Jul 5, 2024

Table of Contents

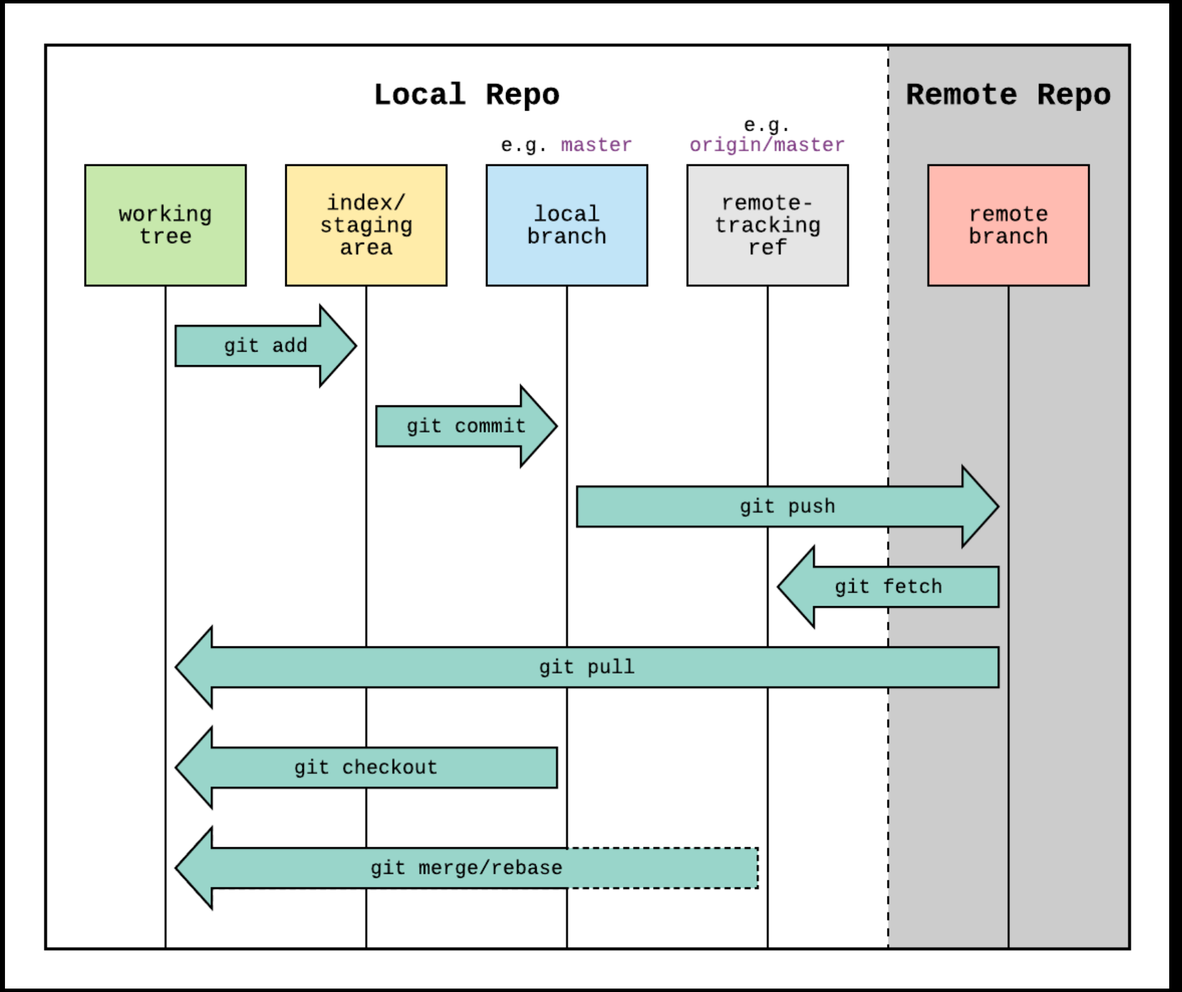
Git & Github

Git

- distributed version control system (VCS)
- tracks changes made to files

Status:

- untracked: file is not tracked/controlled by git
- unmodified: file is the same as the last commit
- modified: file has been changed since the last commit
- staged: file is marked to go into the next commit



Git Configuration

- System Level
    - apply to all users on the system
- Global Level
    - apply to all repositories for the current user
- Local Level
    - apply to the current repository

Git Configuration (cont.)

```
git config --global user.name "rachelie"
git config --global user.email "rachelie.dwg@gmail.com"
git config --global init.defaultBranch "main"
```

- `git config --list` to list all configurations
- `git config --global --unset user.name` to unset a configuration
- `git config --global --edit` to edit configurations
- `git config --global --get user.name` to get a configuration

Basic Commands

- Initializing Repository: `git init`
- Check Repo Status: `git status`
- Adding Files to Staging Area: `git add <file> // . to add all files`
- Commit changes: `git commit -m “message”`
- Commit History: `git log`

Github

www.github.com

- Github는 Git을 이용한 소프트웨어 서비스

- Connecting to Github:

```
git remote add origin <url> # url of the remote repository
git push -u origin main # main is the branch name
```

Clone & Fork

레포를 복사해오기

- Clone: Github에 있는 Repo를 내 로컬로 복사
    - `git clone <url>`

- Fork: 자신의 github 계정에 repo를 복사

.gitignore

git으로 올리고 싶은 내용 선택

- 폴더에 있는 모든 내용은 git으로 올리려고 함
- 관리가 필요하지 않은 파일은 .gitignore에 작성해주면 됨!

이외 내용...

- Branch
- Merge
- Rebase
- Conflict
- Pull Request
- etc…

Crawling Basics

Web Crawling vs. Scraping

- 웹 크롤링
  - URL을 탐색해 반복적으로 링크를 찾아오는 과정
  - 웹페이지를 찾아다니며 정보 수집
  - URL을 수집하고 웹페이지를 복사하여, 수집한 웹페이지에 index 부여
- 웹 스크래핑
  - 우리가 특정한 웹 페이지에서 데이터 추출
  - 특정 웹 존재 ==> 우리가 필요한 정보만 가져옴

Web Crawling vs Scraping

- 웹 크롤링은 특정 웹 페이지를 목표로 하지 않는다
- 일단 탐색 후, 정보를 가져옴 (선탐색 후 추출)

- 웹 스크래핑은 목표로 하는 특정 웹페이지 존재
- 우리가 원하는 정보를 어디서 가져올지 타겟이 분명
- 타겟에서 정보를 가져옴 (‘선결정 후 추출’)

Conda 가상환경 세팅

- 가상환경이 없다면 conda로 create!
- `conda install -c anaconda beautifulsoup4`
- `pip install selenium`
- `pip install requests` ==> 이런식으로 pip로 해도 됨! (이 방법 추천)
- `pip install openpyxl` (excel 변환을 위해)
- pip install lxml

- 그외에 pandas

Developer Tool/Inspection

- right click —> 개발자 도구 창 열기
    - Developer tools, Inspect 등 여러 이름으로 불림
- 원하는 요소를 찾아서 이용

API

우리가 힘드니까~

- Application Programming Interface의 줄임말
- 응용 프로그램에서 사용할 수 있도록 다른 응용 프로그램을 제어할 수 있게 만든 인터페이스
- API를 사용하면 내부 구현 로직을 알지 못해도 정의되어 있는 기능을 쉽게 사용할 수 있음
- 참고: 인터페이스(Interface)란 어떤 장치간 정보를 교환하기 위한 수단이나 방법을 의미함

Blockchain API

https://www.blockchain.com/explorer/api

- json file: https://www.blockchain.com/explorer/api/blockchain_api

감사합니다 :)