

CS2104 Problem Solving in Computer Science

*Margaret Ellis,
Naren Ramakrishnan,
Sehrish Basir Nizamani*



Web Scraping

What is Web Scraping

- An automatic way to retrieve unstructured data from a website and store them in a structured format.

tools and library for web scraping

BeautifulSoup

PYTHON
REQUESTS

Mechanize

Se
Selenium

Scrapy

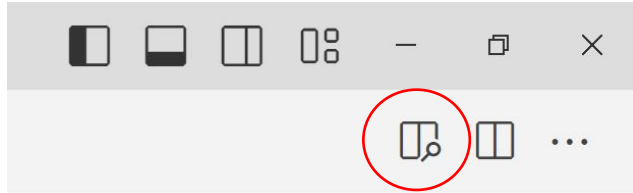
Scraping all up-front websites

1. Inspect the website HTML that you want to crawl
2. Access URL of the website using code and download all the HTML contents on the page
3. Format the downloaded content into a readable format
4. Extract out useful information and save it into a structured format
5. For information displayed on multiple pages of the website, you may need to repeat steps 2–4 to have the complete information.

Scraping websites with API

1. Inspect the XHR network section of the URL that you want to crawl
2. Find out the request-response that gives you the data that you want
3. Depending on the type of request(post or get) and also the request header & payload, simulate the request in your code and retrieve the data from API. Usually, the data got from API is in a pretty neat format.
4. Extract out useful information that you need
5. For API with a limit on query size, you will need to use 'for loop' to repeatedly retrieve all the data

Classwork : reading data off html



Show preview
Side by side view

student_list.html

...

CS2104 > W9 > reading_data_off_html > <code> student_list.html

1

<!doctype html>

2

<html><body><head><title>student

3

<table border="1">

4

<tr>

5

<th>first name</th>

6

<th>last name</th>

7

<th>year</th>

8

</tr> <tr><td>Graham</td><td>Ja

student list

...

← → ↺ http://127.0.0.1:3000/CS2104/W9/readii

≡

This is a student list

sorted by ...

first name	last name	year
Graham	Jarvis	1
Boris	Alvarez	1
Irma	Marquez	1
Gemma	Armstrong	2
Ilina	Mccall	2
Jin	Torres	4
Armand	Delgado	3
Jerry	Small	1



Live Preview v0.4.6 Preview

Microsoft  microsoft.com | 1,443,493 | ★★★★★(47)

Hosts a local server in your workspace for you to preview your ...

Disable Uninstall Switch to Pre-Release Version 

Sample Code:

```
import re
# use regex to parse the html string and extract the target data

in_filename = 'student_list.html'
out_filename = 'student_list.txt'
# defines the name of the file that you're going to read and write into

content = open(in_filename, "r")
output = open(out_filename, "a")
# content and output are file stream object, you could use 'r'(read), 'w'(write) or 'a'(append)

str = content.read()
# read() method will turn a file stream into a regular string

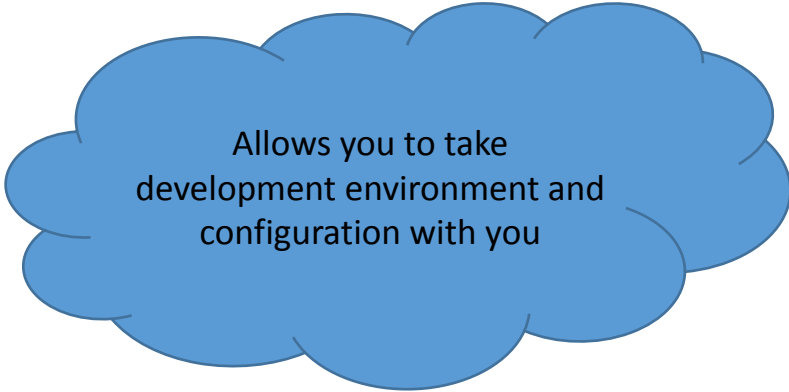
# use matches = re.findall(pattern, str), you need to figure out what pattern to use
# pattern = r'(\w*)</td><td>(\w*)</td><td>(\d) '
# once you get a match, you access different components from the match object using indices.
# use output.write() to write content to the output file
# for output formatting, you can try rjust()

output.close()
content.close()
#close the stream
```


Introduction to Docker



What are containers?



Allows you to take
development environment and
configuration with you

- **Containers** are a method of operating system virtualization that allow you to run an application and its dependencies in resource-isolated processes.
- “A standardized unit of software”
- For example: <https://www.docker.com/why-docker>
- Amazon Web Services, Google Cloud Platform provide this

Why use containers?

Trouble:

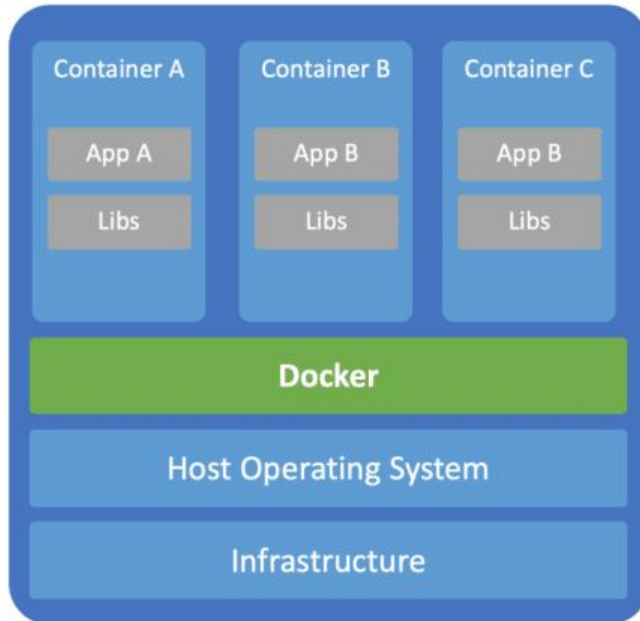
- Path dependency
- Software version
- Environment variables

Benefits:

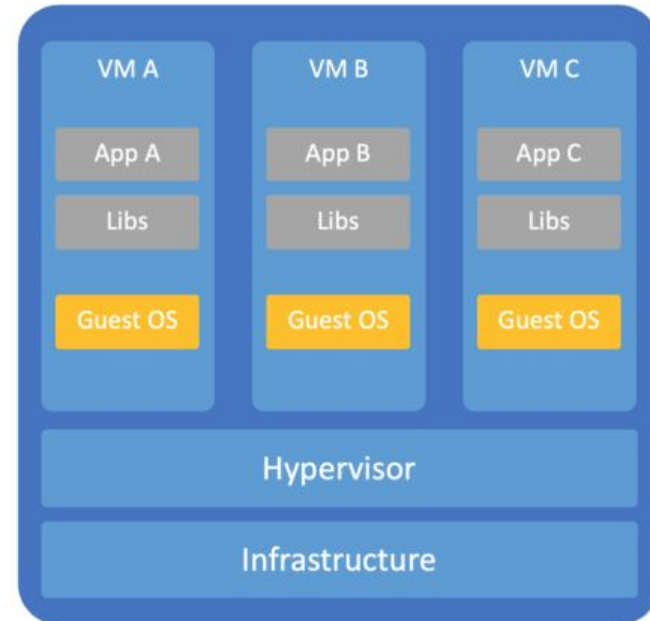
- Encapsulated-portable
- Lightweight-fast
- Demands less resource

Container vs VM

Container

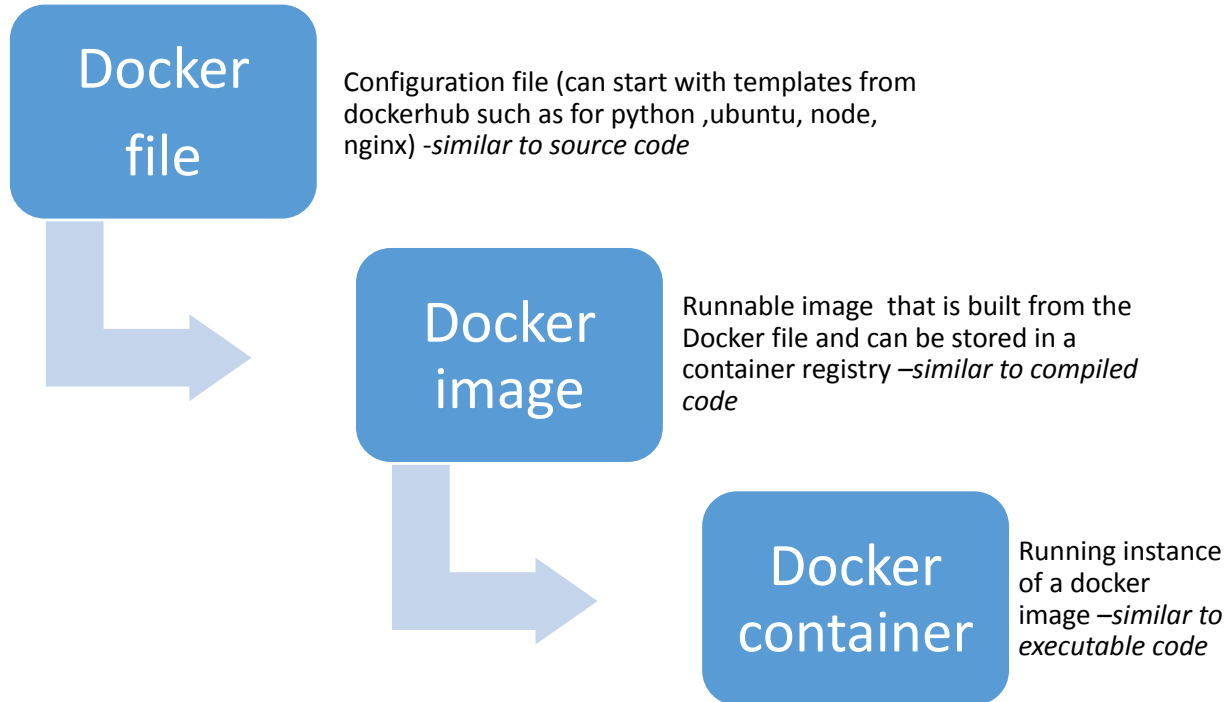


Virtual Machines



Containers

lightweight compared to virtual machines



Classwork: Dockerize a website and deploy it on the cloud

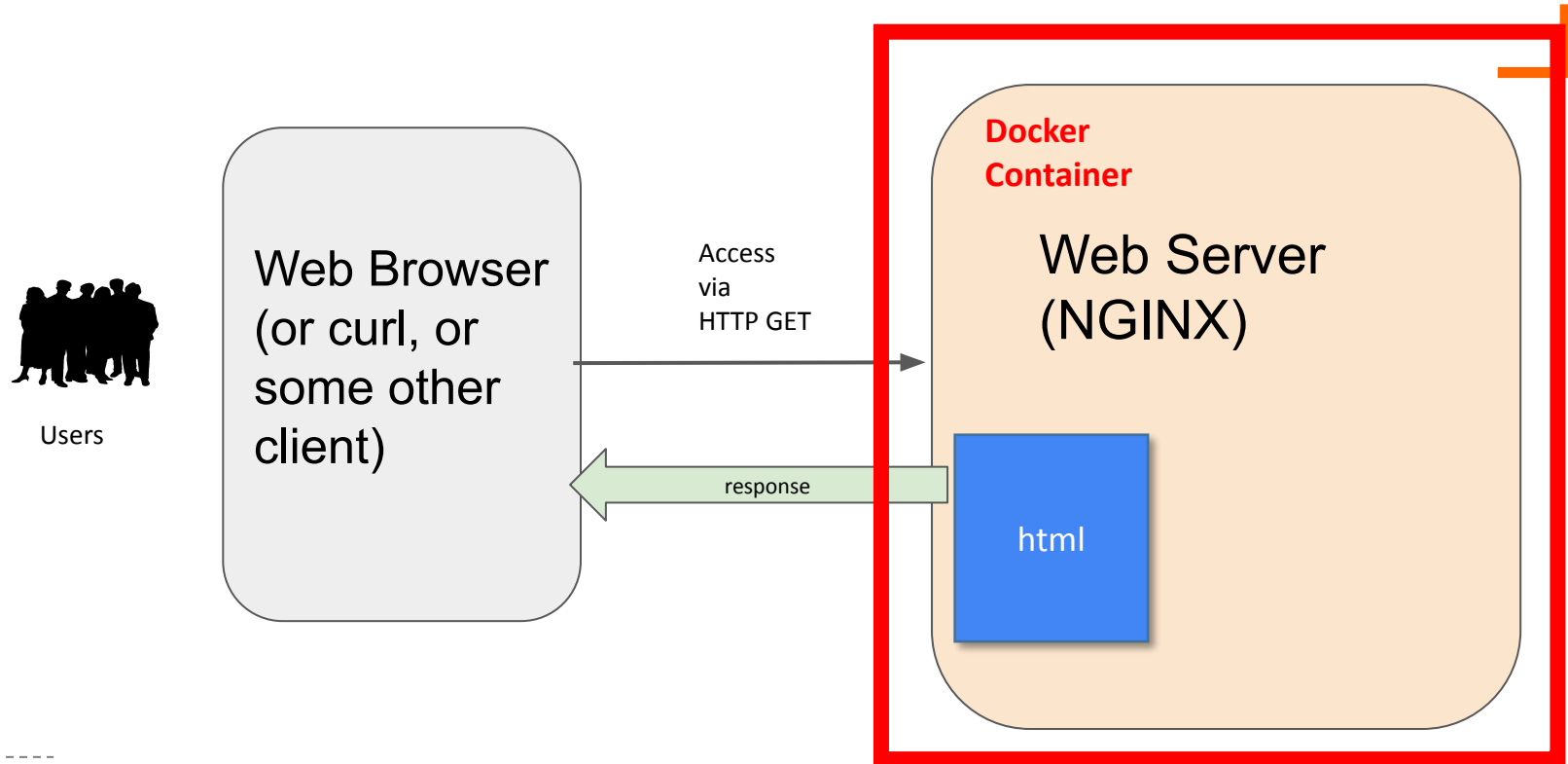
Fork the project on Gitlab

[📄 README](#) [+ Add LICENSE](#) [+ Add CHANGELOG](#) [+ Add CONTRIBUTING](#) [+ Add Kubernetes cluster](#) [+ Set up CI/CD](#)

[⚙️ Configure Integrations](#)

Name	Last commit	Last update
📁 images	Base code	1 week ago
📁 site-content	Base code	1 week ago
🐳 Dockerfile	Base code	1 week ago
📄 README.md	Update instructions to be more specific	1 day ago

Sample Static Web App Architecture



1

Edit the Dockerfile.

Dockerfiles can use another image as a base and then build on top of it, on **rlogin** create a Dockerfile based on nginx:latest

```
FROM nginx:latest  
COPY ...
```

2

Build the docker image.

This takes Dockerfile source code and turns it into a Docker image, which is an executable

3

Push image to **gitlab container registry**. This makes it so your image can be used elsewhere without having to distribute source and rebuild.

Classwork
will guide
you through
these steps!

container.cs.vt.edu

git.cs.vt.edu

container

repository

4

Pull docker image to **Azure cloud** hosting.

5

Run docker image on Azure to host website.

6

Visit your site!

Dockerhub - open source containers

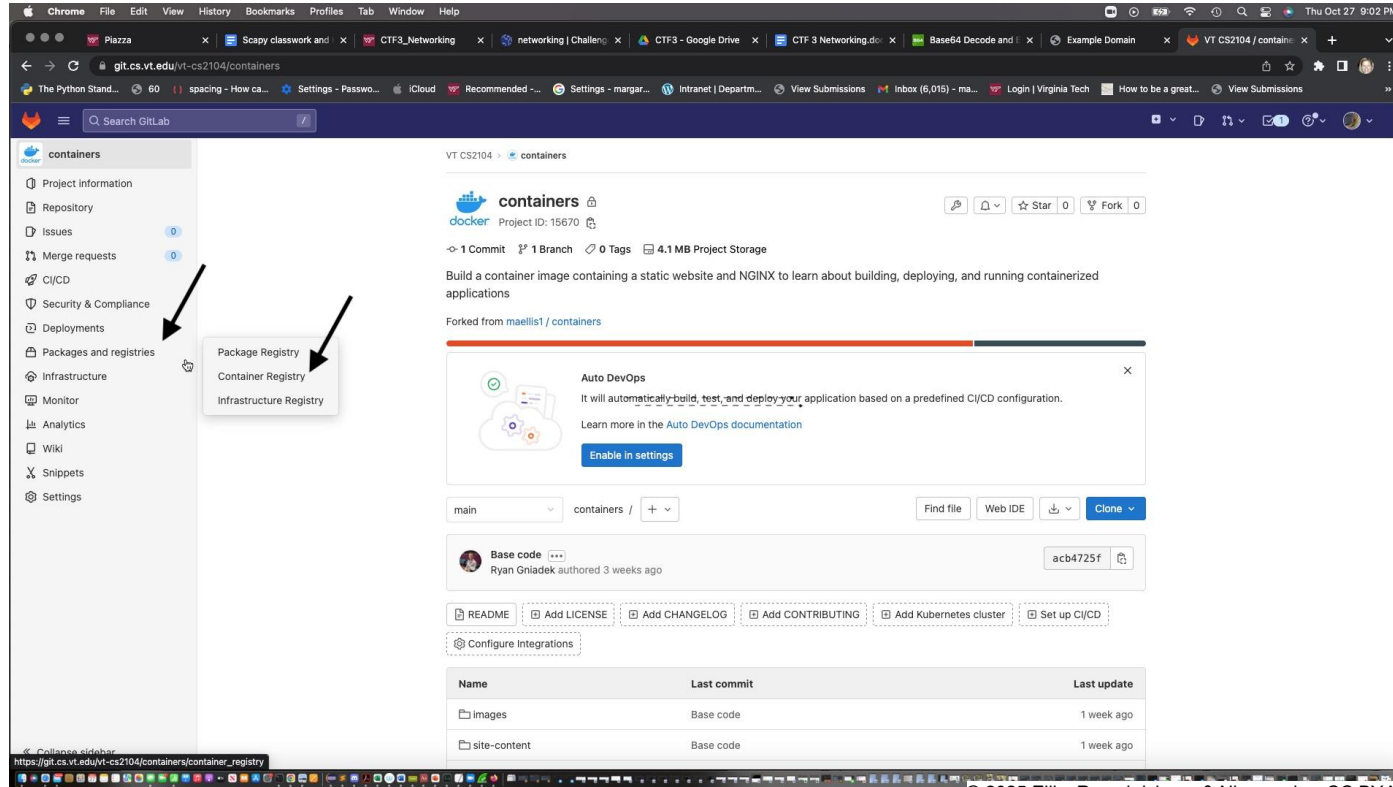
Nginx - web server

Docker image - usable executable

Docker file - source code for Docker image

Docker container - running instance of the Docker image executable

Container registry example in GitLab



The screenshot shows the GitLab web interface for the 'containers' repository. The left sidebar is on the left, and the main content area is on the right. The 'Package registries' section is expanded in the sidebar, and the 'Container Registry' option is highlighted. The main content area shows the repository details for 'containers', including a description, commit history, and a table of images.

Package Registry

- Container Registry
- Infrastructure Registry

containers

Project ID: 15670

1 Commit 1 Branch 0 Tags 4.1 MB Project Storage

Build a container image containing a static website and NGINX to learn about building, deploying, and running containerized applications

Forked from [maellist / containers](#)

Auto DevOps

It will automatically build, test, and deploy your application based on a predefined CI/CD configuration.

Learn more in the [Auto DevOps documentation](#)

[Enable in settings](#)

main containers / +

Find file Web IDE Clone

Base code
Ryan Gniadek authored 3 weeks ago

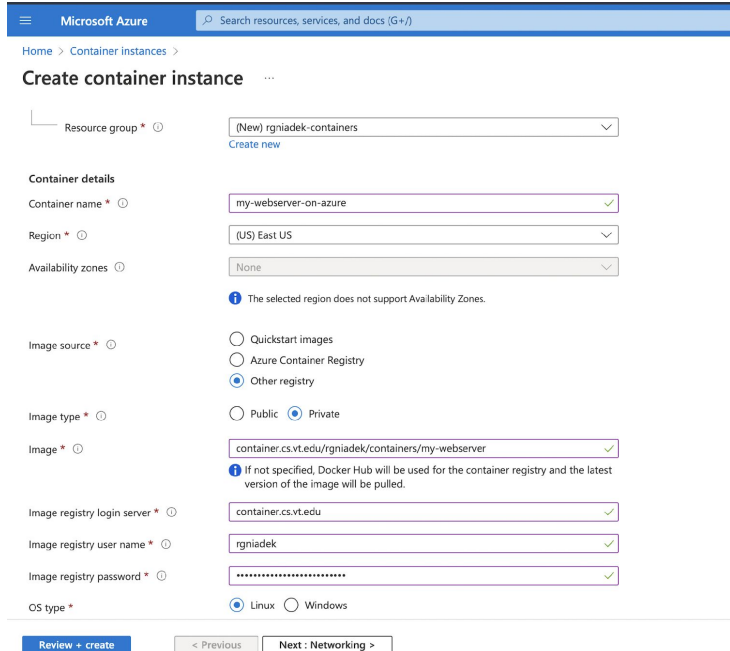
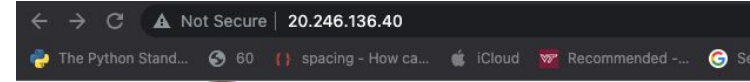
ac4b725f

README Add LICENSE Add CHANGELOG Add CONTRIBUTING Add Kubernetes cluster Set up CI/CD

Configure Integrations

Name	Last commit	Last update
images	Base code	1 week ago
site-content	Base code	1 week ago

Pull container to Azure and host website!

Hello from **your pid's** Nginx container

Edit this in your
index.html using vi
on rlogin