Data Science

Getting Data

June 4th, 2021

This Lecture

Getting some data.

Before we start...

Before we start...

1. Quiz!

Before we start...

1. Quiz!
2. Project 0

Before we start...

1. Quiz!
2. Project 0
3. Course Website

Quiz

Quiz

1. Should be live

Quiz

1. Should be live
2. Should be fixed

Project 0

Project 0

1. Due tonight!

Project 0

1. Due tonight!
2. If you're having issues, let us know.

Course Website

Course Website

1. Many folks email me questions that are answered on the course website.

Course Website

1. Many folks email me questions that are answered on the course website.
2. You should probably check the course website, if it's not answered there, shoot me an email.

Getting Data!

Last week we talked about *data*. Great fine, bit whoop. We want to actually use data!

Python to the rescue

Python to the rescue

There's lots of data on the internet.

Python to the rescue

There's lots of data on the internet. Some of it is useful.

Python to the rescue

There's lots of data on the internet. Some of it is useful. Even less of it is SFW.

Python to the rescue

There's lots of data on the internet. Some of it is useful. Even less of it is SFW. Luckily, we can write programs that let us get this data.

Python to the rescue

How do we do this? A few things to keep in mind:

Python to the rescue

How do we do this? A few things to keep in mind:

1. Every website does their own thing, even if they claim to meet a standard.

Python to the rescue

How do we do this? A few things to keep in mind:

1. Every website does their own thing, even if they claim to meet a standard.
2. You're going to have to get comfortable with exploring what you get back.

Python to the rescue

How do we do this? A few things to keep in mind:

1. Every website does their own thing, even if they claim to meet a standard.
2. You're going to have to get comfortable with exploring what you get back.
3. All hope is not lost, there are some common things that will help.

A light in the darkness

A light in the darkness

1. Learning how JSON works will pay dividends (eventually you won't even think about it)

A light in the darkness

1. Learning how JSON works will pay dividends (eventually you won't even think about it)
2. CSV is crucial and will almost certainly come up.

A light in the darkness

1. Learning how JSON works will pay dividends (eventually you won't even think about it)
2. CSV is crucial and will almost certainly come up.
3. Learning some basic HTML will help, but understand that few sites produce compliant HTML

A light in the darkness

1. Learning how JSON works will pay dividends (eventually you won't even think about it)
2. CSV is crucial and will almost certainly come up.
3. Learning some basic HTML will help, but understand that few sites produce compliant HTML
4. GraphQL is the new kid on the block, unclear how popular it will be (maybe huge!)

Some encouragement

I would be doing you a disservice if I forced you to learn the details of these formats.

Some encouragement

I would be doing you a disservice if I forced you to learn the details of these formats.

1. If you go in thinking that a website has definitely followed the standard, you're only producing tears.

Some encouragement

I would be doing you a disservice if I forced you to learn the details of these formats.

1. If you go in thinking that a website has definitely followed the standard, you're only producing tears.
2. Use an interactive environment (REPL, Jupyter Notebook, etc.)

To the Notebook!

What the title says.

What did we learn?

What did we learn?

1. Be careful with passwords!

What did we learn?

1. Be careful with passwords!
2. Look up the API docs for the website you're trying to use

What did we learn?

1. Be careful with passwords!
2. Look up the API docs for the website you're trying to use
3. Realize the docs are bad

What did we learn?

1. Be careful with passwords!
2. Look up the API docs for the website you're trying to use
3. Realize the docs are bad
4. Go through the stages of acceptance

What did we learn?

1. Be careful with passwords!
2. Look up the API docs for the website you're trying to use
3. Realize the docs are bad
4. Go through the stages of acceptance
5. Explore/play with what you get and press on

What didn't we learn?

What didn't we learn?

1. Beautiful Soup (HTML)

What didn't we learn?

1. Beautiful Soup (HTML)
2. CSV

What didn't we learn?

1. Beautiful Soup (HTML)
2. CSV
3. Manipulating JSON into other formats

What didn't we learn?

1. Beautiful Soup (HTML)
2. CSV
3. Manipulating JSON into other formats
4. For all of these: read the docs of the libraries!

Thanks for your time!