python™

Intro to Python

A Crash Course in Programming

school of ai

Startberry

Who am I?

Marcin Kraszewski, developer/analyst/content editor at *Devskiller*, co-founder of *Improv.pl*. Dean of the *Warsaw School of AI*. Host of *the Yellow Duck* Podcast.

What is Python?

Python is an **interpreted** high-level **programming language** for general-purpose programming. Created by Guido van Rossum and first released in **1991**, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Who uses Python?

Why should I learn Python if I’m not a Software Engineer?

- Automate boring tasks
- Help Manage a complex personal project
- Develop a tool for your family/friends
- Learn other subjects while learning programming

Isn’t Excel/Google Docs Good Enough?

Python is for much more than just number crunching. It’s capable of doing all of the following:

- Web development
- Data analysis
- Task Automation
- Data Collection
- Data Processing
- And more!

The Zen of Python

```
Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one-- and preferably only one --obvious way to do it.
Although that way may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than *right* now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking great idea -- let's do more of those!
```

Live Coding #1 Python Basics

Google Colab

http://colab.research.google.com

What is Web Scraping?

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.[1] Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to **automated processes implemented using a bot or web crawler**. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

What is a REST API?

Representational State Transfer (REST) is an architectural style that defines a set of constraints to be used for creating web services. Web services that conform to the REST architectural style, or RESTful web services, provide interoperability between computer systems on the Internet. REST-compliant web services allow the requesting systems to access and manipulate textual representations of web resources by using a uniform and predefined set of stateless operations.

If API's exist why would I waste the time to write my own scraper?

If there is data that you’d like to get access to and available via the website but not via the API one of the only alternatives is to use write a web scraper or use a service that will scrape the data for you.

Actual Web Scraping Businesses

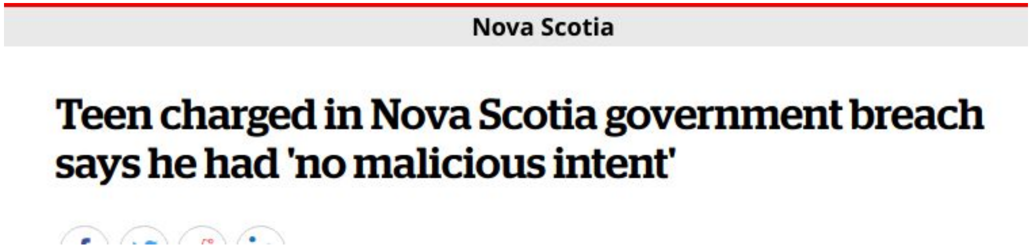
- Collect ALL of the company job post pages in the world and sell access to the data. (vacancysoft.com)
- Collect the mugshots of people from all of the police databases in the USA and offer to remove them if the person in the photo pays to have them removed (MUGSHOTS.com)
- Collect all the worlds data and let people search for it. (GOOGLE)

Ethics of Web Scraping

- Collect only the information that you **NEED**
- Don’t ignore **Robots.txt**!
- Try to run your scrapers at times that are lower usage times for the website
- Only write a web scraper if you need the information and the API is too limited

Dangers of Web Scraping

- Man sued by Facebook for trying to build a service around collecting Facebook data: https://petewarden.com/2010/04/05/how-i-got-sued-by-facebook/



Nova Scotia

Teen charged in Nova Scotia government breach says he had 'no malicious intent'

- Teen in Nova Scotia arrested for downloading ‘accidently’ publicly posted government data.

What is a library?

A library is also a collection of implementations of behavior, written in terms of a language, that has a well-defined interface by which the behavior is invoked. For instance, people who want to write a higher level program can use a library to make system calls instead of implementing those system calls over and over again

How do I use a library in my project?

- Visit Github.com
- Find a library that interests you
- Read the docs
- Install the library
- Import the library
- Done!

What are the pros and cons of using a library vs. writing a solution from scratch?

Although there are many great libraries out there make sure to do the following due diligence before using a library in a commercial or serious project:

- Read the license
- See when the last commit was made
- Make sure there are valid (up-to-date) well written docs

What is Python requests?

HTTP for Humans

http://docs.python-requests.org/en/master/

Live Coding #2 Requests Intro

Google Colab

http://colab.research.google.com

What is BeautifulSoup?

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Live Coding #3 BeautifulSoup Intro

Google Colab

http://colab.research.google.com

Next Steps

Take what you've learned tonight and create your own customized webscraping application to collect data that interests you.

* Use the **Flask** web framework to create a simple admin panel to easily control your scraping application. The application should have the following features:

* Set the script schedule (every 12 hours, every 24 hours, every week, or once per month)

* Set the base_url that the app will use when scraping

Post your code to Github so that we can discuss it at the next meeting. The 2 best projects (in terms of code quality and execution) will be featured on our Facebook page and mentioned at the next meeting.