NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp

Web Scraping Project

Introduction

Alright -- so you've pretty much mastered visual representations of your data. Your client is so impressed with your ability to lead them through the big picture and is very pleased that you can deliver insights seamlessly to the global community. What's next? The client is beginning to see all the value you bring to the company, but they want more. They're ready for you to do some original research for them. You have free reign on the topic, as long as you tell them something interesting and useful -- great! But there's a catch -- they haven't given you any data. What are you supposed to do? Well, you still need to tell a story, but now you've got to do the leg work on the data collection yourself. Once again, what story will you tell?

What We're Looking For

You're now becoming a multilingual programmer -- both R and Python skills are being fostered simultaneously -- and the bootcamp doesn't show any signs of slowing down. Machine learning topics are whizzing by, homeworks are getting more complicated, and sometimes you can't remember that Python begins indexing at 0 instead of 1 like in R. How are you going to survive?

The bottom line is: **You will find a way.**

For this project, your primary task is to collect data from a web source by method of scraping. What you do with that data after its collection is up to you (e.g., numeric description, basic/interactive graphics, machine learning, etc.); however, you still must lead the audience through an overall insight. While it is required you scrape your data using Python, the analyses following are language agnostic -- but remember that the

primary task is to foster data scraping skills. Use of an API can be beneficial in practice -- but they don't always exist. A creative application of scraping may allow the collection of web data that isn't readily downloadable or available elsewhere. What does this mean? Now you have the additional task of data collection on top of munging, visualization, insight gathering, and storytelling -- but the same amount of time. How will this all be possible?

The bottom line is: **You will still find a way.**

As always, preparation will be key. Successful projects will encompass a plethora of skills including, but not limited to, the following:

- Submission in respect to the deadline.
- Background knowledge of dataset(s).
- Communication of motivation: why do we care?
- Research questions of interest: what do you want to find out?
- Answers to research questions: what have you uncovered?
- Presentation skills.
- Time management (not going over the allotted time).
- Ability to answer audience questions effectively and efficiently.
- Balance of complexity and simplicity.
- Explanation of future work: what would you do if given more time, data, etc.?
- Demonstration of web scraping ability in Python (without relying solely on API calls).
- Additional use of R (e.g., ggplot2, Shiny, etc.), Python (e.g., NumPy, SciPy, Pandas, Matplotlib, Seaborn, etc.), and/or machine learning techniques.

The Details

Your project proposal declaration is due uniformly by the beginning of the Pulse Check on **Friday, November 4**. **No exceptions.** You must declare your dataset on the project proposal document and give a short background on some initial research questions. A

sentence or two will suffice. These may change completely as you proceed with your analysis -- this is ok.

This is an **individual project** in respect to the final deliverable. **No exceptions.** Every student must have their own project and presentation; however, please feel free to collaborate and help each other with coding problems, insights, brainstorming, etc. We welcome this!

All code, data, etc. used to generate your graphics and/or Shiny app and any slides, markdown files, etc. intended for your presentation are due to the project GitHub repository uniformly by **Sunday, November 13 at 11:59pm**. **No exceptions.**

You will be required to deliver a **10 minute presentation** and respond to any audience questions. Time slots will be randomly assigned on this calendar, so all projects must be submitted on time. **No exceptions.**

An associated blog post will be due by **Sunday, November 20 at 11:59pm**. **No exceptions.** Remember, this is a living and breathing document. You may continue to develop and edit your project far beyond the deadline, as no project will ever truly be complete.

For inspiration, take a look at our previous students' blog posts (here's a link specifically to the Web Scraping category).

For any lingering questions, please do not hesitate to reach out; we are always here to help!

Good luck!