

Lab 11: Web Scraping

CSE/IT 107

NMT Computer Science

“The sooner you start to code, the longer the program will take.”

— R. Carlson

“Any sufficiently advanced technology is indistinguishable from magic.”

— A. Clarke

“Imagination is more important than knowledge.”

— A. Einstein

1 Introduction

1.1 HTML

As you may be familiar, HTML is the main formatting language of the Internet.

1.2 BeautifulSoup

What do?

How do?

We gave it to you.

Example

1.3 Matplotlib

What do?

How do?

1.4 Wikipedia

That one thing that russell found (about the philosophy link)

2 Exercises

Boilerplate

Remember that this lab *must* use the boilerplate syntax introduced in Lab 5.

`file.py`

3 Submitting

Files to submit:

- file.py (Section 2)

You may submit your code as either a tarball (instructions below) or as a .zip file. Either one should contain all files used in the exercises for this lab. The submitted file should be named either `cse107_firstname_lastname_lab11.zip` or `cse107_firstname_lastname_lab11.tar.gz` depending on which method you used.

For Windows, use a tool you like to create a .zip file. The TCC computers should have 7z installed. For Linux, look at lab 1 for instructions on how to create a tarball or use the “Archive Manager” graphical tool.

Upload your tarball or .zip file to Canvas.