WEB SCRAPING
STEP-BY-STEP CHECKLIST

STEP 1:
DETERMINE THE PROBLEM YOU WANT TO SOLVE

- ☐ Do you want to build up a database?
- ☐ Do you want to perform analytics later?
- ☐ Do you want to build an application reliant on live data?

STEP 2:
BRAINSTORM THE CONCRETE SOLUTION

- ☐ What data type are you scraping, is it JSON, HTML, csv files or something else?
- ☐ If HTML, do you need to scrap static or dynamic HTML? Is it reliant on live data?

STEP 3:
FIND YOUR GOOD, SOLID DATA SOURCE

- ☐ Does my source provide accurate data?
- ☐ Is my source up-to-date?
- ☐ Does my source provide large amount of data?

STEP 4:
VIEW PAGE SOURCE

Chrome: Customize → More Tools → Developer Tools
Safari: Right Click + Show Page Source
Firefox: Press Alt + Tools → Web Developer → Page Source
Microsoft Edge: More Icon → Developer Tools
Internet Explorer: Press Alt + View → Source

STEP 5:
LOOK AT HTML

```
<table> This shows the data below in a table format
<tr> This is the first row
<th> This is the header of the first column </th>
<th> This is the header of the second column </th>
</tr>
<tr> This is the second row
<td> First table item in second row </td>
<td> Second table item in second row </td>
</tr>
</table>
```

STEP 6:
CREATE AN ALGORITHM

For the example to the left, our algorithm is going to read everything between <table> and </table>. Our data is stored between the <td></td> tags, so our algorithm will save everything inside of those tags. Grabbing the headers is optional.

STEP 7:
SCRAPE YOUR DATA FROM THE SITE

- ☐ Run your code to pull the website page source as text and the algorithm will pull out exactly what you specified before.
- ☐ Double check your data formatting to find bugs that you have missed before.

STEP 8:
STRUCTURE YOUR DATA

- ☐ Do you want to work with pandas later? Maybe consider JSON to make creating data frames easy
- ☐ Do you want to make your data easily sharable? Consider a csv file. Maybe you want to save into a SQL or NoSQL database?

STEP 9:
SAVE + READ YOUR DATA

- ☐ Save your data into a database or file
- ☐ Make sure the file is easily accessible to you and there are no naming conflicts
- ☐ Read your data in the format that you created

STEP 10:
SET NOTIFICATIONS

- ☐ Set trigger values - if you're streaming live data, maybe set trigger values that trigger events once certain values are detected
- ☐ If your trigger is triggered, how are you going to be made aware? Maybe you want to send an email or a text to your device, or maybe you want to automatically post to twitter?

STEP 11:
SET UP A TIMER

You don't always want to be clicking run on your script.

- ☐ Set up a timer on your script so that it runs when you start up your computer, open your Mail, connect to your wifi, etc..
- ☐ Or... upload the script to a web service, and having it running constantly

STEP 12:
USE YOUR DATA

- ☐ Use your data to solve your problem

That's basically it - you did it. Great job! Now go forth and scrape all the data you can find!

CODING WITH MAX