13 – SCRAPING SINGLES

Websites can have a lot of useful information on them. Sometimes though our program might need to get this information live. The only issue with doing this is that python will download the whole thing as a HTML string that would then need to be filtered through. Luckily for you, this is the task.

Step 1: Read the website into python

This code is pretty standard. The only thing different from the manual is the link that is being read. In this case, the link is: https://www.officialcharts.com/charts/singles-chart/

```
import urllib.request

fp = urllib.request.urlopen("https://www.officialcharts.com/charts/singles-chart/")
mybytes = fp.read()

mystr = mybytes.decode("utf8")
fp.close()
```

Step 2: Understand the HTML

When you right click on a page and press inspect, you can view the code that created the page. Using this information, you can look for patterns that can be taken advantage of. This webpage has the following two tags that I am interested in:

```
▼<td>
    <span class="position">1</span>
  </td>
▶<td>…</td>
▼<td>
  ▼<div class="track">
    ▶<div class="cover" style="height:60px;width:60px">…
    </div>
    ▼<div class="title-artist">
      ▼<div class="title">
          <a href="/search/singles/funky-friday/">FUNKY
          FRIDAY</a> == $0
        </div>
```

**** and **<a href=”**. Each of these would allow me to splice the string down so that it just shows the position of the track in the chart and then the name of the track.

Step 3: Create the main loop

Ultimately what you have stored in the variable **mystr** is a massive string. This can be spliced using **mystr**[:]. How do we know where to splice it though? To search through **mystr** you would write **a = mystr.find(“search here”)** which would either return the index where the string “search here” started or -1 if not found. You could even write another search that starts after where a was found with this **b = mystr.find(“search here”, a+1).** So back to the main loop, well **while a > -1:** would be enough to get you started. Remember to do another check later to keep a changing e.g. **a = mystr.find(“search here”, a+1)**.