Web Crawler
Intro to Web Science
10 Points

You are to write a web crawler in Python which starts with a seed URL and downloads and saves web pages to disk. The crawler should only save HTML resources and ignore other content types.

Start with the Python crawler from the PowerPoint presentation at

http://www.harding.edu/fmccown/classes/comp475-s13/Python-Web-Crawler.pptx

and make the following modifications:

1. Add an optional parameter *limit* with a default of 10 to `crawl()` function which is the maximum number of web pages to download,

2. Save files to *pages* dir using the MD5 hash of the page's URL. Example:

```
import hashlib
filename = 'pages/'  + hashlib.md5(url.encode()).hexdigest() + '.html'
```

3. Only crawl URLs that are in harding.edu domain (*.harding.edu). Use a regular expression when examining discovered links. Regex example:

```
import re
 p = re.compile('ab*')
 if p.match('abc'):
    print("yes")
```

Turn in your crawler.py program to Easel before class on the due date.