

Learn Log - Exeter MicroApp Coding Challenge:

Pavitra Mohandas - SVCE

Day-1: (Friday, 30th Oct 2020)

1. Looked into the basics of web scrapping. Read through various articles in Medium.
2. Once understanding the basics, I was able to understand the requirements better and was initially looking through how urls are extracted from a given website.
3. I came across various packages like LinkGrabber , Urllib and BeautifulSoup for performing the operation.
4. Having looked into LinkGrabber, I was able to see that along with link other attributes such as Class, id, text, seo were returned.

Output:

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

Pavitras-MacBook-Air:exeter pavitra$ python3 file.py
[ { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.kriyadocs.com/',
  'id': 'jupiter-custom',
  'seo': '',
  'text': 'Kriyadocs'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/services/',
  'id': 'jupiter-custom',
  'seo': '',
  'text': ''},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/services/#Editorial_Services',
  'id': 'jupiter-custom',
  'seo': '#Editorial_Services',
  'text': 'EDITORIAL SERVICES'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/services/#Data_Services',
  'id': 'jupiter-custom',
  'seo': '#Data_Services',
  'text': 'DATA SERVICES'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/services/#Artwork',
  'id': 'jupiter-custom',
  'seo': '#Artwork',
  'text': 'ARTWORK & DESIGN'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/services/#Project_Management',
  'id': 'jupiter-custom',
  'seo': '#Project_Management',
  'text': 'PROJECT MANAGEMENT'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': '#',
  'id': 'jupiter-custom',
  'seo': '#',
  'text': '#'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/blog/',
  'id': 'jupiter-custom',
  'seo': '',
  'text': 'BLOG'},
  { 'class': ['menu-item-link', 'js-smooth-scroll'],
  'href': 'https://www.exeterpremedia.com/about-us/',
  'id': 'jupiter-custom',
  'seo': ''}
```

5. Later, I looked into what internal and external links are in a website. I understood the concept and it was easy to correlate
6. Urllib's specific packages such as urlparse, urljoin for finding internal, external links.
7. I proceeded to code and I had to was able to arrive at the output that I desired
8. Being a visual learner, I was automatically drawn to Colorama package in python which will help us distinguish between internal and external links with various colors.
9. I decided to fix Green Color for Internal Link and Blue Color for External Link.
10. I tested it with two different websites,
 1. <https://www.exeterpremedia.com/services/>
 2. <https://www.coursera.org/programs/sri-venkateswara-college-yqrql>

Output:

```
External link: https://www.kriyadocs.com/
Internal link: https://www.exeterpremedia.com/services/
Internal link: https://www.exeterpremedia.com/blog/
Internal link: https://www.exeterpremedia.com/about-us/
Internal link: https://www.exeterpremedia.com/events/
Internal link: https://www.exeterpremedia.com/news/
Internal link: https://www.exeterpremedia.com/contact-us/
External link: https://careers.exeterpremedia.com/jobs/Careers
Internal link: https://www.exeterpremedia.com/
Internal link: https://www.exeterpremedia.com
Internal link: https://www.exeterpremedia.com/client/
Internal link: https://www.exeterpremedia.com/exeter-privacy-policy/
Internal link: https://www.exeterpremedia.com/terms-and-conditions/
External link: tel://+91-44-23452922
External link: tel://+44-20-3287-2783
External link: tel://+1-646-736-7767
External link: mailto://sales@exeterpremedia.com
External link: https://twitter.com/Exeter_Premedia
External link: https://www.linkedin.com/company/exeter-premedia-services
External link: https://www.facebook.com/exeterpremediaservices
External link: https://www.instagram.com/exeterpremedia/
```

Work completed Today:

1. Getting all the links for a specific website.
2. Checking if a given URL is valid.
3. Classifying links in a given website as Internal or External.

Day-2:(Saturday, 31st October 2020)

1. Looked on how to identify titles from a given link. Was able to get title of the given URL only.
2. Took a very long time to check how the titles can be obtained for Top Level Pages.

Output:

```
Title: Exeter - Privacy Policy - Exeter
Title: Exeter - All about authors and publishers
Title: Client - Exeter
Title: Exeter - Contact Us
Title: Exeter - Events
Title: Exeter - Services
Title: Exeter - News
Title: Exeter - Blog
Title: Exeter - Happy Authors and Delighted Publishers
Title: Terms and Conditions - Exeter
```

3. Looked up online if any package was available to get homepage if a link is given as an input. Was unable to find any.
4. Derived the logic for printing the home page of a link if any other URL rather than home page is given.

Output:

```
Home Page Link: https://www.exeterpremedia.com
```

5. Understood what a meta tag is. Was able to write code for getting {key, value} pairs of meta name and content.

6. The number of internal links, external links, and total was calculated at the end of the day.

Output:

```
Total External links: 10  
Total Internal links: 11  
Total links: 21
```

7. Looked into how to obtain image urls from a given website. Discovered the package extraction. Found the image link and image size along
8. Understood the logic of website size by looking through various websites. Found the size of an entire website.

Word Completed Today:

1. Page titles of all the top level pages obtained
2. Homepage given any link was derived
3. Meta Name and Meta Content values obtained
4. Total number of internal, external links obtained
5. Image URL from websites obtained
6. Website size obtained

Day-3:(Sunday, 1st November 2020)

1. Looked online to check for any packages to obtain entire website data. Came to know BeautifulSoup had a function called stripped_strings. Was able to get data in the form of list. Converted it into string.
2. Understood what stop words are and passed the string to stopwords to remove stop words present in the website
3. Classified unigram and bigram from the non-stop words data.
4. Found the 20 top frequencies for unigram and bigram. Arranged them in descending order.

- Exported values such as Home Page Link, External Link, Internal Link, Title, Total External links, Total Internal links, Website Size, Data from website, Parsed Text to text file.
- Exported values such as Meta_Name, Meta_Content, Unigram, Bigram, Top 20 Unigram and Bigram, Image Url

Output:

CSV File:

Meta_Name	Meta_Content	Unigram	Top 20 Frequency Unigram	Bigram	Top 20 Bigram	Image URL
0	viewport	Enter	(", 100)	(Enter, ", 1)	(", 100), 10	https://www.kontagprmedia.com/wp-content/uploads/2018/05/kontagprmedia-8pg-entip
1	format-detection	-	(", 40)	(", "Services)	(", "Ow", 0)	https://www.kontagprmedia.com/wp-content/uploads/2018/05/16650734104232046-1.png
2	description	Services	(Services, 10)	(Services, "Kiyadoc)	(Enter, ", 7)	
3	bettercast	Enter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	Kiyadoc	(services, ")	(Editorial, Services), 0	
4	betterdescription	en_US	services	(Rer, 10)	(Services, EDITORIAL)	(Data, Services), 0
5	bettertitle	article	EDITORIAL	(", 10)	(EDITORIAL, SERVICES)	(Project, Management), 0
6	bettertable	Enter - Services	SERVICES	(", 10)	(SERVICES, DATA)	(T, ", 0)
7	bettercreator	Enter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	DATA	(", 10)	(DATA, SERVICES)	(Artwork,), 0
8	generator	https://www.kontagprmedia.com/services/	SERVICES	(", 10)	(SERVICES, ARTWORK)	(R, Design), 0
9	generator	Enter	ARTWORK	(Design, 10)	(ARTWORK, R)	(T, ", 0)
10	generator	summary_large_image	A	(publishers, 10)	(R, DESIGN)	(T, Address), 0
11	generator	Enter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	DESIGN	(Ow, 0)	(DESIGN, PROJECT)	(authors, publishers), 0
12	generator	Enter - Services	PROJECT	(Enter, 0)	(PROJECT, MANAGEMENT)	(turnaround, times), 0
13	reapplication-Thumbnail	Enter, Premedia	MANAGEMENT	(content, 0)	(MANAGEMENT, Resource)	(turnaround, times), 0
14	generator	Enter, Premedia	Resources	(R, 7)	(Resource, BLOG)	(Vib, ", 0)
15		Powered by LayerSlider 6.7.6 - Multi-Purpose, Responsive, Parallax, Mobile-Friendly Slider Plugin for WordPress.	BLOG	(authors, 7)	(BLOG, About)	(", Including), 0
16		WordPress 5.2	About	(books, 7)	(About, us)	(page, composition), 0
17		MasterSlider 3.2.7 - Responsive Touch Image Slider	us	(US, 0)	(us, ABOUT)	(Design, ", 0)
18		user	ABOUT	(Date, 0)	(ABOUT, US)	(Design, services), 0
19		May 9, 2019	US	(Artwork, 0)	(US, EVENTS)	(books, journals), 0
20		June 3, 2019	EVENTS		(EVENTS, NEWS)	
21		Enter	NEWS		(NEWS, CONTACT)	
22		Powered by WPBakery Page Builder - drag and drop page builder for WordPress.	CONTACT		(CONTACT, US)	
23		Powered by Slider Revolution 5.4.8 - responsive, Mobile-Friendly Slider Plugin for WordPress with comfortable drag and drop interface.	US		(US, careers)	
24		https://www.kontagprmedia.com/wp-content/uploads/2018/05/cropped-efvicon-1.png.pagespeed.ic_.._jlk2gucl-273d73.png	careers		(careers, Editorial)	
25		Juniper 6.1.1	Kiyadoc		(Kiyadoc, "services)	
26			services		(services, EDITORIAL)	
27			EDITORIAL		(EDITORIAL, SERVICES)	
28			SERVICES		(SERVICES, DATA)	
29			DATA		(DATA, SERVICES)	
30			SERVICES		(SERVICES, ARTWORK)	
31			ARTWORK		(ARTWORK, R)	
32			A		(R, DESIGN)	
33			DESIGN		(DESIGN, PROJECT)	
34			PROJECT		(PROJECT, MANAGEMENT)	
35			MANAGEMENT		(MANAGEMENT, Resource)	
36			Resources		(Resource, BLOG)	
37			BLOG		(BLOG, About)	
38			About		(About, us)	
39			us		(us, ABOUT)	
40			ABOUT		(ABOUT, US)	
41			US		(US, EVENTS)	
42			EVENTS		(EVENTS, NEWS)	
43			NEWS		(NEWS, CONTACT)	
44			CONTACT		(CONTACT, US)	
45			US		(US, careers)	
46			careers		(careers, Editorial)	
47			Editorial		(Editorial, Services)	
48			Services		(Services, Data)	
49			Data		(Data, Services)	
50			Services		(Services, Artwork)	
51			Artwork		(Artwork, R)	
52			A		(R, DESIGN)	
53			Design		(Design, Project)	
54			Project		(Project, Management)	
55			Management		(Management, Editorial)	
56			Editorial		(Editorial, Services)	
57			Services		(Services, Data)	
58			Data		(Data, Services)	
59			Services		(Services, Artwork)	
60			Artwork		(Artwork, R)	
61			A		(R, DESIGN)	
62			Design		(Design, Project)	
63			Project		(Project, Management)	
64			Management		(Management, Enter)	
65			Enter		(Enter, ", 1)	
66			"		(", and-to-end)	
67			and-to-end		(and-to-end, "signal)	
68			digital		(digital, publishing)	
69			publishing		(publishing, services)	
70			services		(services, designed)	
71			designed		(designed, enable)	
72			enable		(enable, authors)	
73			authors		(authors, publishers)	
74			publishers		(publishers, obtain)	

Home Page Link: <https://www.exeterpremedia.com>

```

External Link: https://www.krityado.co.s/
Internal Link: https://www.xorterpremedia.com/servlets/
Internal Link: https://www.xorterpremedia.com/blog/
Internal Link: https://www.xorterpremedia.com/about-us/
Internal Link: https://www.xorterpremedia.com/event/
Internal Link: https://www.xorterpremedia.com/news/
Internal Link: https://www.xorterpremedia.com/contact-us/
External Link: https://careers.xorterpremedia.com/about-careers
Internal Link: https://www.xorterpremedia.com/
Internal Link: https://www.xorterpremedia.com/terms/
Internal Link: https://www.xorterpremedia.com/client-privacy-policy/
Internal Link: https://www.xorterpremedia.com/about-us-and-conditions/
External Link: tel://+91-44-23452522
External Link: tel://+91-28-3287-2703
External Link: tel://+91-44-246-746-717
External Link: mailto://sales@xorterpremedia.com
External Link: https://twitter.com/Xorterpmedia
External Link: https://www.linkedin.com/company/xorterpmedia-services
External Link: https://www.facebook.com/xorterpmedia/
External Link: https://www.instagram.com/xorterpmedia/

```

Title: Exorter - Privacy Policy - Exorter
Title: Exorter - All about authors and publishers
Title: Client - Exorter
Title: Exorter - Contact Us
Title: Exorter - Events
Title: Exorter - Services
Title: Exorter - News
Title: Exorter - Blog
Title: Exorter - Happy Authors and Delighted Publishers
Title: Terms and Conditions - Exorter

Total External links: 10
Total Internal links: 11
Total links: 21

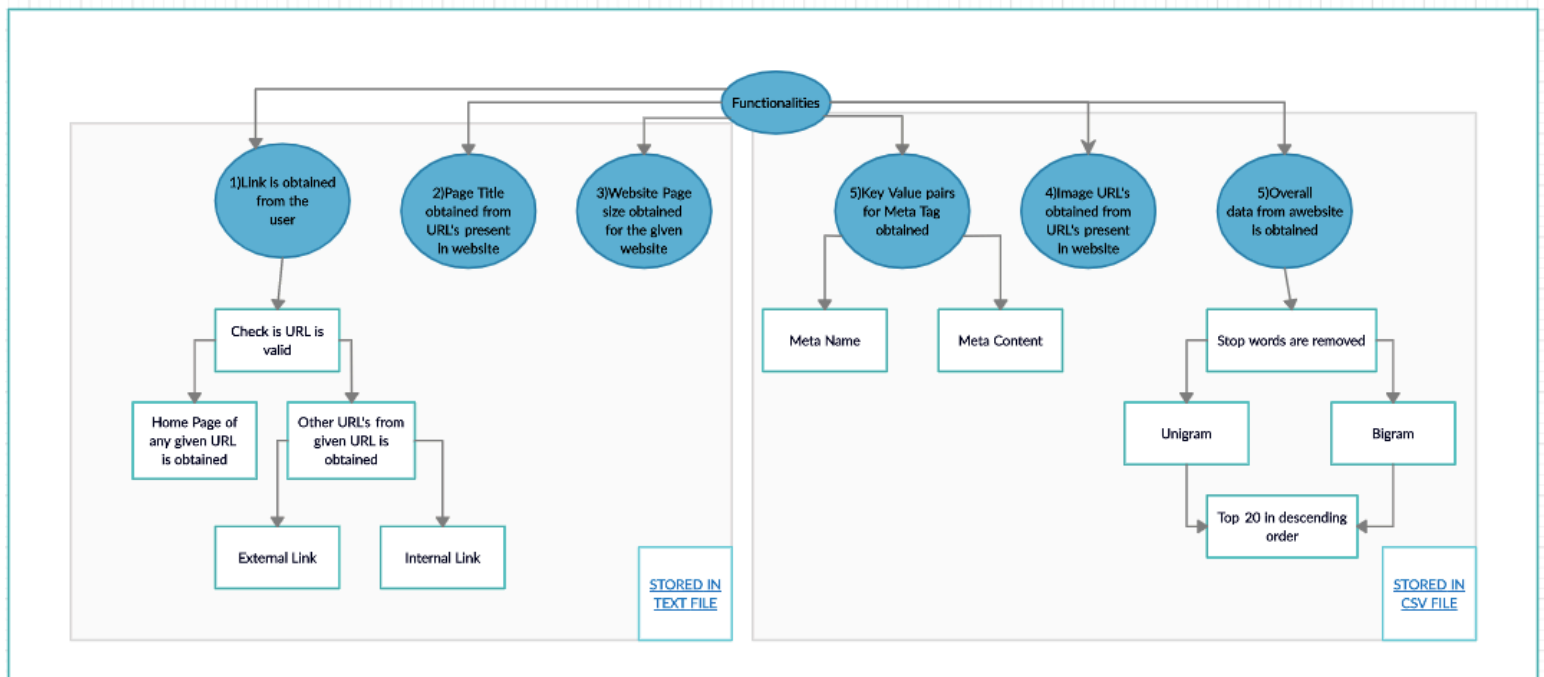
Website Size: 126.919921875 mb

[illegible][illegible]

Major Problems I encountered:

1. Setting up virtual environment was a pretty hard task. It took sometime to setup the Virtual environment for python.
2. Every data stored into variables was of different shape. Hence I initially faced a lot of errors while writing data to the CSV file. Was able to resolve the problem when I used ignore_text in pandas.
3. A variable called 'bi' in my code was a generator. So while passing the data to the CSV file, I was facing errors. Resolved it by converting the generator to a list.
4. Accessing title from all the url's present in a website was a very tedious process.

Block Diagram:



In the above block diagram, I have explained all the functionalities which is available in the program that I have developed. The output file is stored in two formats based on the functionality.

Decisions Taken:

1. Initially used linkGrabber package for extracting links. But I was unable to perform operation for finding internal and external link. So I had to scrape it and then use Urllib library for performing the task of extracting all the urls.
2. Decided to have two different files to induce more readability to the code. Establishing certain data in the form of tables made it easy to understand.
3. Made sure to have a function for checking validity of URL since there are high chances for user to enter a value which might not be a valid URL.

Learning:

1. Understood the concepts related to web scrapping.
2. Learnt the packages involved in web scrapping.
3. Realized how python plays a very important role in development. It makes things really easy and the packages help save a lot of memory.
4. Various tags involved in web development.

Summary:

Personally, I was very happy that I got an opportunity to be a part of the MicroApp Challenge since, I was able to effectively use 52 hours in the last 2.5 days for solving a problem. I became very curious while performing the task and got very involved into researching stuff related to Web Scrapping. I got very excited when I was getting desired outputs for the given modules. Having attended 24 hour hackathons in the past, this challenge gave me the confidence that I will be able to attend 60 hour hackathons in future. After a very long time, my boredom came to an end and I was very delighted that I was able to finish the task which was given to me.