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:

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 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.
5. 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.
6. Exported values such as Meta_Name, Meta_Content, Unigram, Bigraam, Top 20 Unigram and Bigram, Image Url

Final Flow of Modules:

User enters URL -> Validation of URL is done -> Home page of URL obtained -> All the URL's inside website is obtained -> Internal External URL's identified -> Page Titles Obtained -> Image URL obtained -> Meta Keyvalue pairs obtained -> Count of links in website -> Website size -> Data from page obtained -> Stop words removed -> With parsed text, unigram bigram obtained -> Top 20 obtained -> Data exported to CSV and Text Files

Output:—CSV File:

	Meta_Name	Meta_Content	Unigram	Top 20 Frequency Unigram	Bigram	Top 20 Bigram	Image URL
0	viewport	width=device-width, initial-scale=1.0, minimum-scale=1.0, maximum-scale=1.0, user-scalable=0	Exeter	('.', 103)	('Exeter', '-')	('.', 'We'), 10)	https://www.exeterpremedia.com/wp-content/uploads/2019/08/exeterpremedia-logo-.webp
1	format-detection	IE=edge,chrome=1	-	('.', 46)	('-', 'Services')	('.', 'Our'), 8)	https://www.exeterpremedia.com/wp-content/uploads/2019/08/1158577581042333696-1.png
2	description	telephone=no	Services	('Services', 15)	('Services', 'Kriyadocs')	('Exeter', ''), 7)	
3	twitter:card	Exeter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	Kriyadocs	('services', 15)	('Kriyadocs', 'services')	('Editorial', 'Services'), 5)	
4	twitter:description	en_US	services	('We', 15)	('services', 'EDITORIAL')	('Data', 'Services'), 3)	
5	twitter:title	article	EDITORIAL	('(', 15)	('EDITORIAL', 'SERVICES')	('Project', 'Management'), 3)	
6	twitter:site	Exeter - Services	SERVICES	('-', 14)	('SERVICES', 'DATA')	('(', ')'), 3)	
7	twitter:creator	Exeter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	DATA	('(', 14)	('DATA', 'SERVICES')	('Artwork', '&'), 4)	
8	generator	https://www.exeterpremedia.com/services/	SERVICES	('', 12)	('SERVICES', 'ARTWORK')	('&', 'Design'), 4)	
9	generator	Exeter	ARTWORK	('Design', 10)	('ARTWORK', '&')	('(', ')'), 4)	
10	generator	summary_large_image	&	('publishers', 10)	('&', 'DESIGN')	('(', 'Adobe'), 4)	
11	generator	Exeter provides editorial services such copyediting, proofreading and indexing, XML conversion services, artwork services and project management services.	DESIGN	('Our', 9)	('DESIGN', 'PROJECT')	('authors', 'publishers'), 3)	
12	generator	Exeter - Services	PROJECT	('Exeter', 8)	('PROJECT', 'MANAGEMENT')	('turnaround', 'times'), 3)	
13	msapplication-TileImage	@Exeter_Premedia	MANAGEMENT	('content', 8)	('MANAGEMENT', 'Resources')	('times', ','), 3)	
14	generator	@Exeter_Premedia	Resources	('&', 7)	('Resources', 'BLOG')	('etc', ','), 3)	
15		Powered by LayerSlider 6.7.6 - Multi-Purpose, Responsive, Parallax, Mobile-Friendly Slider Plugin for WordPress.	BLOG	('authors', 7)	('BLOG', 'About')	('(', 'including'), 3)	
16		WordPress 5.2	About	('books', 7)	('About', 'us')	('page', 'composition'), 3)	
17		MasterSlider 3.2.7 - Responsive Touch Image Slider	us	('US', 6)	('us', 'ABOUT')	('InDesign', ','), 3)	
18		user	ABOUT	('Data', 6)	('ABOUT', 'US')	('Design', 'services'), 3)	
19		May 9, 2019	US	('Artwork', 6)	('US', 'EVENTS')	('books', 'journals'), 3)	
20		June 3, 2019	EVENTS		('EVENTS', 'NEWS')		
21		Exeter	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.exeterpremedia.com/wp-content/uploads/2019/05/cropped-sfavicon-1.png.pagespeed.ic._z8c2gbut3-270x270.png	careers		('careers', 'Kriyadocs')		
25		Jupiter 6.1.1	Kriyadocs		('Kriyadocs', 'services')		
26			services		('services', 'EDITORIAL')		
27			EDITORIAL		('EDITORIAL', 'SERVICES')		
28			SERVICES		('SERVICES', 'DATA')		
29			DATA		('DATA', 'SERVICES')		
30			SERVICES		('SERVICES', 'ARTWORK')		
31			ARTWORK		('ARTWORK', '&')		
32			&		('&', 'DESIGN')		
33			DESIGN		('DESIGN', 'PROJECT')		
34			PROJECT		('PROJECT', 'MANAGEMENT')		
35			MANAGEMENT		('MANAGEMENT', 'Resources')		
36			Resources		('Resources', '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', '&')		
52			&		('&', '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', '&')		
61			&		('&', 'Design')		
62			Design		('Design', 'Project')		
63			Project		('Project', 'Management')		
64			Management		('Management', 'Exeter')		
65			Exeter		('Exeter', '')		
66			'		('', 'end-to-end')		
67			end-to-end		('end-to-end', 'digital')		
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')		

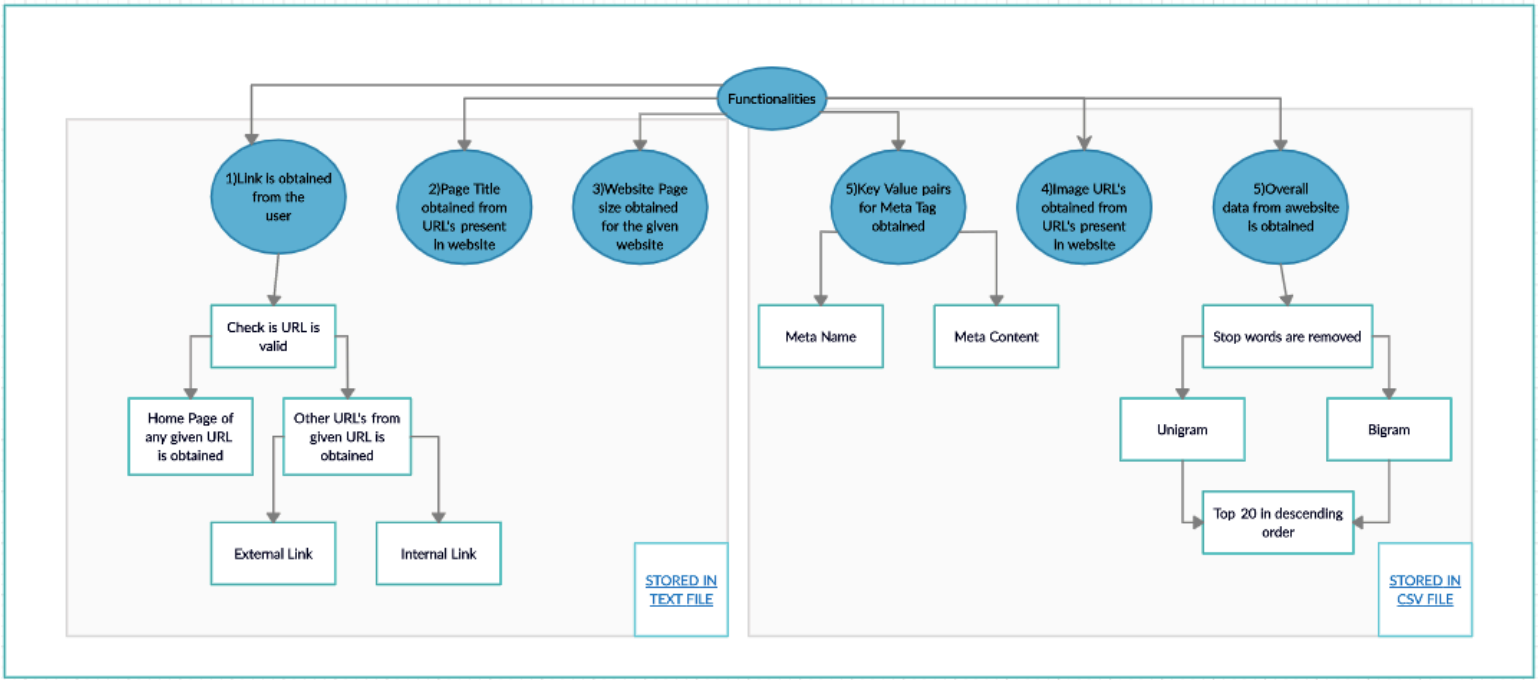
Output — Text File:

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


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/
Internal 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://+44-20-3287-2703
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/




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


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


Website Size: 126.919921875 mb


Data from website:
Exeter - Services Kriyadocs services EDITORIAL SERVICES DATA SERVICES ARTWORK & DESIGN PROJECT MANAGEMENT Resources BLOG About us ABOUT US EVENTS NEWS CONTACT US careers Kriyadocs services EDITORIAL SERVICES DATA SERVICES ARTWORK & DESIGN PROJECT MANAGEMENT Resources BLOG About us ABOUT US EVENTS NEWS CONTACT US careers Editorial Services Data Services Artwork & Design Project Management Editorial Services Data Services Artwork & Design Project Management Exeter's end-to-end digital publishing services are designed to enable authors and publishers obtain output of outstanding quality with minimal turnaround times. Our processes are streamlined and platform-driven, leaving very little room for unintended manual errors. We have dedicated teams with specialized skill sets in defined service categories copyediting and proofreading. They are experienced in producing publications in the varied fields of Science, Technical, Medicine (STM), education, social sciences etc. covering the breadth of academic and scholarly publishing, and continually keep themselves updated on changes in standards and technology advances. Our services spectrum includes Editorial Services, Data Services, Artwork & Design and end-to-end Project Management. Editorial Services Exeter's Editorial Services can help you achieve editorial excellence, so that you can stand out and succeed in your publishing goals. We are experts in the business of publishing content and have perfected the art and science of editorial nuances which meets the needs of publishers and authors looking for highest quality and quick turnaround times. Machine learning underlies our editorial processes - thus ensuring high degree of quality and constantly evolving capability. Copyediting We can turn your good book into a great book through workflow-driven, collaborative and comprehensive copyediting services. We give primacy to retaining the author's voice and understanding the context through direct interactions with authors. Machine learning-driven examination of the text for flow, grammar, errors, repetitions, style mismatch or any other inconsistencies ensures a fine copy. The checks provided include - Spelling, including UK and US English Grammar, including rephrasing sentences References and footnotes for the style of the publishing, technical accuracy and placement of citations Appropriate and consistent captioning for figures and tables Mathematical equations for subscripts, superscripts, exponents and symbols Proofreading Proofreading is perhaps the most important step in producing a ready-for-publication book or journal. Exeter's highly proficient team ensures that this crucial check of the copyedited text produces output that is consistent, complies with the publisher's house style and meets the highest standards. Apart from checking for typical errors, in style, fonts, capitalization, italicization, hyphenation, correct usage of symbols (for example in equations), we also ensure consistency of typefaces and hierarchies. Be it journal articles, books, PhD theses, dissertations, or proposals in the STM field - we proofread and produce polished content. Indexing An index is not a collection of mere keywords or phrases, but an organized map of contents, including cross-references, grouping of like concepts and other useful information for intellectual analysis (especially, for non-trade publications). The absence of it may make even a good subject content go unnoticed by the reader. Exeter's professional indexers ensure your content is indexed comprehensively with related concepts in different sections of the book being cross-referenced or grouped. We can create simple, keyword based, flat (single-level) author and subject indexes to complex, topical, multi-level indexes for content-intensive books or other publications that increase the value of your publication manifold. Data Services Exeter's Data Services help publishers go digital. Backed by more than 15 years of experience, our data conversion services enable you to not only create new publication but also migrate your existing print versions to multiple digital formats. We can be your single point of contact and can ensure successful content conversions from various data sources in any typeset format as needed by you without compromising on quality. Typesetting Our typesetting and page composition services are based on XML-first workflow that preserves the integrity of the XML file throughout the various stages of the production cycle ensuring that the XML file is compliant to the publisher's DTD at any stage, which means that the XML can be generated literally on demand. These end-to-end page composition services catering to books and STM journals range from simple text pages to those containing complex mathematical equations; single column to multi-column and single colored to multi-colored presentation. We are well-versed in JATS, InDesign, LaTeX, QuarkXPress and XSL-FO. Data Conversion We can convert any composite material such as equations and tables, from any page composition software (such as HTML, Adobe InDesign, QuarkXpress, MS Word) or even hard copies to XML, HTML, MS Word, PDF and ebook formats (PDF, ePUB, Mobipocket, KF8, Nook etc). Our innovative, systematic best-practice based data capture, archival imaging, metadata creation, DTD/schema development to XML transformation means our data conversion services are of high standards, delivered time wise while meeting all your unique requirements. We provide ebook conversion, DTBook conversion - Digital Accessible Information System (DAISY) and National Instructional Materials Accessibility Standard (NIMAS), XML Conversion, PubMed Conversion Services and more. This includes enhanced ePub conversions with reflow of text as well as audio- and video-enabled, HTML5-based interactive eBooks. Artwork & Design Exeter's Artwork and Design services enable publishers create visually pleasing, dynamic art layouts that enhance content value and learning outcomes. From intricate design templates to simple vector redraws, our creative team has the experience and tools to transform your ideas into dynamic designs. Our Artwork and Design services include comprehensive illustrations, graphic designs, websites and multimedia designs for a broad range of books and journals. Illustration Services We process figures, illustrations and redraw artwork based on a rough sketch provided by the author. Our team has extensive experience in graphics software including Adobe Photoshop, Adobe Illustrator and Corel Draw. We also handle color scanning and color correction. Our illustrations and artwork services include Situational illustrations, Technical and Medical illustrations and Illustrations for children's, fiction, non-fiction books. Design Services We have a team of professional designers in all capacities with extensive experience in Adobe Photoshop, Adobe Illustrator, Adobe InDesign, Corel Draw, and QuarkXPress who can render artwork and designs to meet the requirements of a wide range of clients from conventional STM/Academic publishers, to publishers of school books/children's books/educational material etc. Our Design services portfolio includes Graphic Design covering Books, Covers (CD or magazine covers), Infographics and marketing collateral; and Multimedia Design (HTML5 based interactive e-Books, Audio and Video editing, Voice-overs to scripts, 2D / 3D graphics creation and animation). End-to-End Project Management Exeter's Project Management services ensure that authors and publishers reap the benefits of our global best practices, proven technologies, skilled professionals and faster turnaround times. We manage end-to-end production of books and journals for print and digital media (multiple online and mobile formats). We partner with authors and publishers, take complete ownership to ensure that their raw manuscripts are turned into finely-polished and ready-for-publication content. Our comprehensive project management services cover: Analyzing and understanding the needs of publishers Working together with publishers, editors and proofreaders Interacting directly with authors for query resolutions, seeking copyright permissions for material (figures, tables, text) being reproduced from other published sources), etc. Managing all stages of the digital content production lifecycle of the scholarly publication from copyediting, proofreading, typesetting, consolidation of revisions received from authors, editors, proofreaders to a common "master proof" to generation of the final electronic file Tracking, reporting, communication, schedule management and efficient resource management The complete project lifecycle from manuscript to the final book or journal is workflow-based (with distinct workflows for books and journals) ensuring efficient turnaround time and cost-effectiveness. SiteMap - Home - Services - Clients - Blog - Careers - About Us - News - Events - Contact Us - Privacy Policy - Terms & Conditions - ISO/IEC 27001 : 2013 - ISO 9001 : 2015 Address INDIA (Corporate Office) Ascendas International Tech Park Unit-3, Level-1, Pinnacle CSIR Road, Taramani Chennai 600113 Tamil Nadu, India India: +91-44-23452922 UK: +44-20-3287-2703 USA: +1-646-736-7767 sales@exeterpremedia.com FOLLOW US ON Location Copyright All Rights Reserved © 2019 Contact Us We're not around right now. But you can send us an email and we'll get back to you, asap. Send


Parsed Text:
Exeter - Services Kriyadocs services EDITORIAL SERVICES DATA SERVICES ARTWORK & DESIGN PROJECT MANAGEMENT Resources BLOG About us ABOUT US EVENTS NEWS CONTACT US careers Kriyadocs services EDITORIAL SERVICES DATA SERVICES ARTWORK & DESIGN PROJECT MANAGEMENT Resources BLOG About us ABOUT US EVENTS NEWS CONTACT US careers Editorial Services Data Services Artwork & Design Project Management Editorial Services Data Services Artwork & Design Project Management Exeter ' end-to-end digital publishing services designed enable authors publishers obtain output outstanding quality minimal turnaround times . Our processes streamlined platform-driven , leaving very little room unintended manual errors . We dedicated teams specialized skill sets defined service categories copyediting proofreading . They experienced producing publications varied fields Science , Technical , Medicine ( STM ) , education , social sciences etc , covering breadth academic scholarly publishing , continually keep updated changes standards technology advances . Our services spectrum includes Editorial Services , Data Services , Artwork & Design end-to-end Project Management . Editorial Services Exeter ' Editorial Services help achieve editorial excellence , stand succeed publishing goals . We experts business publishing content perfected art science editorial nuances meets needs publishers authors looking highest quality quick turnaround times . Machine learning underlies editorial processes - thus ensuring high degree quality constantly evolving capability . Copyediting We turn good book great book workflow-driven , collaborative comprehensive copyediting services . We give primacy retaining author ' voice understanding context direct interactions authors . Machine learning-driven examination text flow , grammar , errors , repetitions , style mismatch inconsistencies ensures fine copy . The checks provided include - Spelling , including UK US English Grammar , including rephrasing sentences References footnotes , style publishing , technical accuracy placement citations Appropriate consistent captioning figures tables Mathematical equations subscripts , superscripts , exponents symbols Proofreading Proofreading perhaps important step producing ready-for-publication book journal . Exeter ' highly proficient team ensures crucial check copyedited text produces output consistent , complies publisher ' house style meets highest standards . Apart checking typical errors style , fonts , capitalization , italicization , hyphenation , correct usage symbols ( example equations ) , also ensure consistency typefaces hierarchies . Be journal articles , books , PhD theses , dissertations , proposals STM field - proofread produce polished content . Indexing An index collection mere keywords phrases , organized map contents , including cross-references , grouping like concepts useful information intellectual analysis ( especially , non-trade publications ) . The absence may make even good subject content go unnoticed reader . Exeter ' professional indexers ensure content indexed comprehensively related concepts different sections book cross-referenced grouped . We create simple , keyword based , flat ( single-level ) author subject indexes complex , topical , multi-level indexes content-intensive books publications increase value publication manifold . Data Services Exeter ' Data Services help publishers go digital . Backed 15 years experience , data conversion services enable create new publication also migrate existing print versions multiple digital formats . We single point contact ensure successful content conversions various data sources typeset format needed without compromising quality . Typesetting Our typesetting page composition services based XML-first workflow preserves integrity XML file throughout various stages production cycle ensuring XML file compliant publisher ' DTD stage , means XML generated literally demand . These end-to-end page composition services catering books STM journals range simple text pages containing complex mathematical equations ; single column multi-column single colored multi-colored presentation . We well-versed JATS , InDesign , LaTeX , QuarkXPress XSL-FO . Data Conversion We convert composite material equations tables , page composition software ( HTML , Adobe InDesign , QuarkXpress , MS Word ) even hard copies XML , HTML , MS Word , PDF ebook formats ( PDF , ePUB , Mobipocket , KF8 , Nook etc ) . Our innovative , systematic best-practice based data capture , archival imaging , metadata creation , DTD/schema development XML transformation means data conversion services high standards , delivered time meeting unique requirements . We provide ebook conversion , DTBook conversion - Digital Accessible Information System ( DAISY ) National Instructional Materials Accessibility Standard ( NIMAS ) , XML Conversion , PubMed Conversion Services . This includes enhanced ePub conversions reflow text well audio- video-enabled , HTML5-based interactive eBooks . Artwork & Design Exeter ' Artwork Design services enable publishers create visually pleasing , dynamic art layouts enhance content value learning outcomes . From intricate design templates simple vector redraws , creative team experience tools transform ideas dynamic designs . Our Artwork Design services include comprehensive illustrations , graphic designs , websites multimedia designs broad range books journals , Illustration Services We process figures , illustrations redraw artwork based rough sketch provided author . Our team extensive experience graphics software including Adobe Photoshop , Adobe Illustrator Corel Draw . We also handle color scanning color correction . Our illustrations artwork services include Situational illustrations , Technical Medical illustrations Illustrations children ' fiction , non-fiction books . Design Services We team professional designers capacities extensive experience Adobe Photoshop , Adobe Illustrator , Adobe InDesign , Corel Draw , QuarkXPress render artwork designs meet requirements wide range clients conventional STM/Academic publishers , publishers school books/children ' books/educational material etc . Our Design services portfolio includes Graphic Design covering Books , Covers ( CD magazine covers ) , Infographics marketing collateral ; Multimedia Design ( HTML5 based interactive e-Books , Audio Video editing , Voice-overs scripts , 2D / 3D graphics creation animation ) . End-to-End Project Management Exeter ' Project Management services ensure authors publishers reap benefits global best practices , proven technologies , skilled professionals faster turnaround times . We manage end-to-end production books journals print digital media ( multiple online mobile formats ) . We partner authors publishers , take complete ownership ensure raw manuscripts turned finely-polished ready-for-publication content . Our comprehensive project management services cover : Analyzing understanding needs publishers Working together publishers , editors proofreaders Interacting directly authors query resolutions , seeking copyright permissions material ( figures , tables , text ) reproduced published sources ) , etc . Managing stages digital content production lifecycle scholarly publication copyediting , proofreading , typesetting , consolidation revisions received authors , editors , proofreaders common " master proof " generation final electronic file Tracking , reporting , communication , schedule management efficient resource management The complete project lifecycle manuscript final book journal workflow-based ( distinct workflows books journals ) ensuring efficient turnaround time cost-effectiveness . SiteMap - Home - Services - Clients - Blog - Careers - About Us - News - Events - Contact Us - Privacy Policy - Terms & Conditions - ISO/IEC 27001 : 2013 - ISO 9001 : 2015 Address INDIA ( Corporate Office ) Ascendas International Tech Park Unit-3 , Level-1 , Pinnacle CSIR Road , Taramani Chennai 600113 Tamil Nadu , India India : +91-44-23452922 UK : +44-20-3287-2703 USA : +1-646-736-7767 sales @ exeterpremedia.com FOLLOW US ON Location Copyright All Rights Reserved © 2019 Contact Us We 're around right . But send us email 'll get back , asap . Send
```

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 hackathones 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.