**Data Collection Assignment #1 – UN Country Offices Websites**

The `AI (Artificial Intelligence) Powered Knowledge Base Prototype Development` project requires SDG (Sustainable Development Goals) labelled text data that is collected from UN websites’ blogposts, news, publications, and others. Furthermore, data collection from social media platforms may be considered in the future.

3 different resources have been identified for the data collection process and each resource will be considered as an assignment. The assignment is the data collection process by using Python and notebooks. You may need to create functions that scrape specific web URLs, scape URLs over specific URL patterns, at the end you would come up with the texts on the scraped from the websites.

The assignment #1 is scraping the UN Country Offices (COs) Websites. Please carefully read the below.

|  |  |
| --- | --- |
| A sample country office website | <https://www.jo.undp.org/> |

Country office websites have the same URL path structure as:

|  |  |  |
| --- | --- | --- |
| https://www. | COUNTRY\_CODE | .undp.org/ |

Since the country office websites have been switching to another type of design, some of the COs websites do not provide desired pages anymore. Desired pages can be found in:

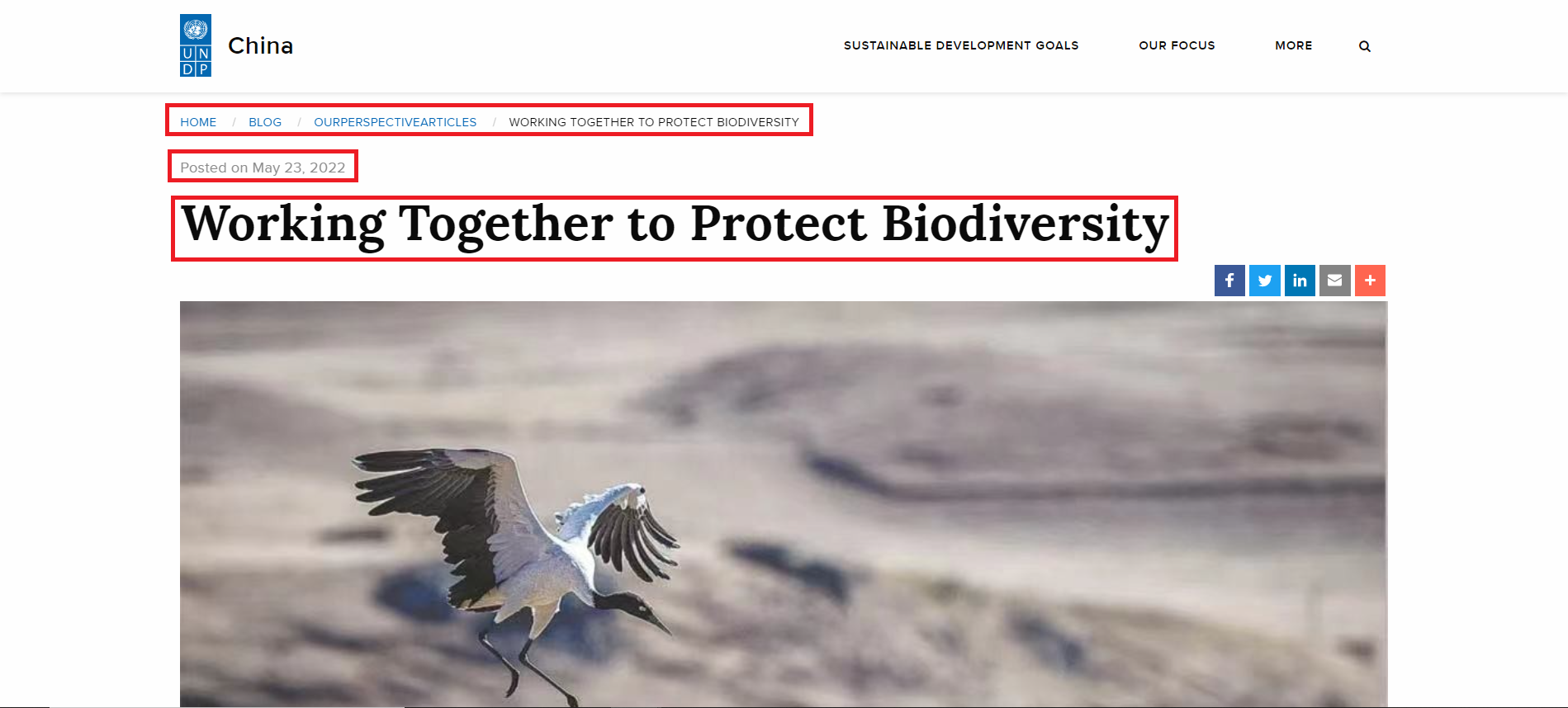
|  |  |  |
| --- | --- | --- |
| Page naming | URL Path Pattern | Image |
| Sustainable Development Goals Inquiry Page (sdg 16) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/sustainable-development-goals/goal-16-peace-justice-and-strong-institutions.html |  |
| Learn More Page (the path is specifically for sdg 16) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/learnmore.html?tagid=sdg-advisors%3Asdg16-peace-justice-and-strong-institutions |  |
| Blog Page (all) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/blog.html |  |
| News Centre (all) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/news-centre.html |  |
| Stories Page (all) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/stories.html |  |
| Press center (all) | https://www. COUNTRY\_CODE.undp.org/content/ COUNTRY\_NAME(OR SPECIFIC NAME)/en/home/presscenter.html |  |

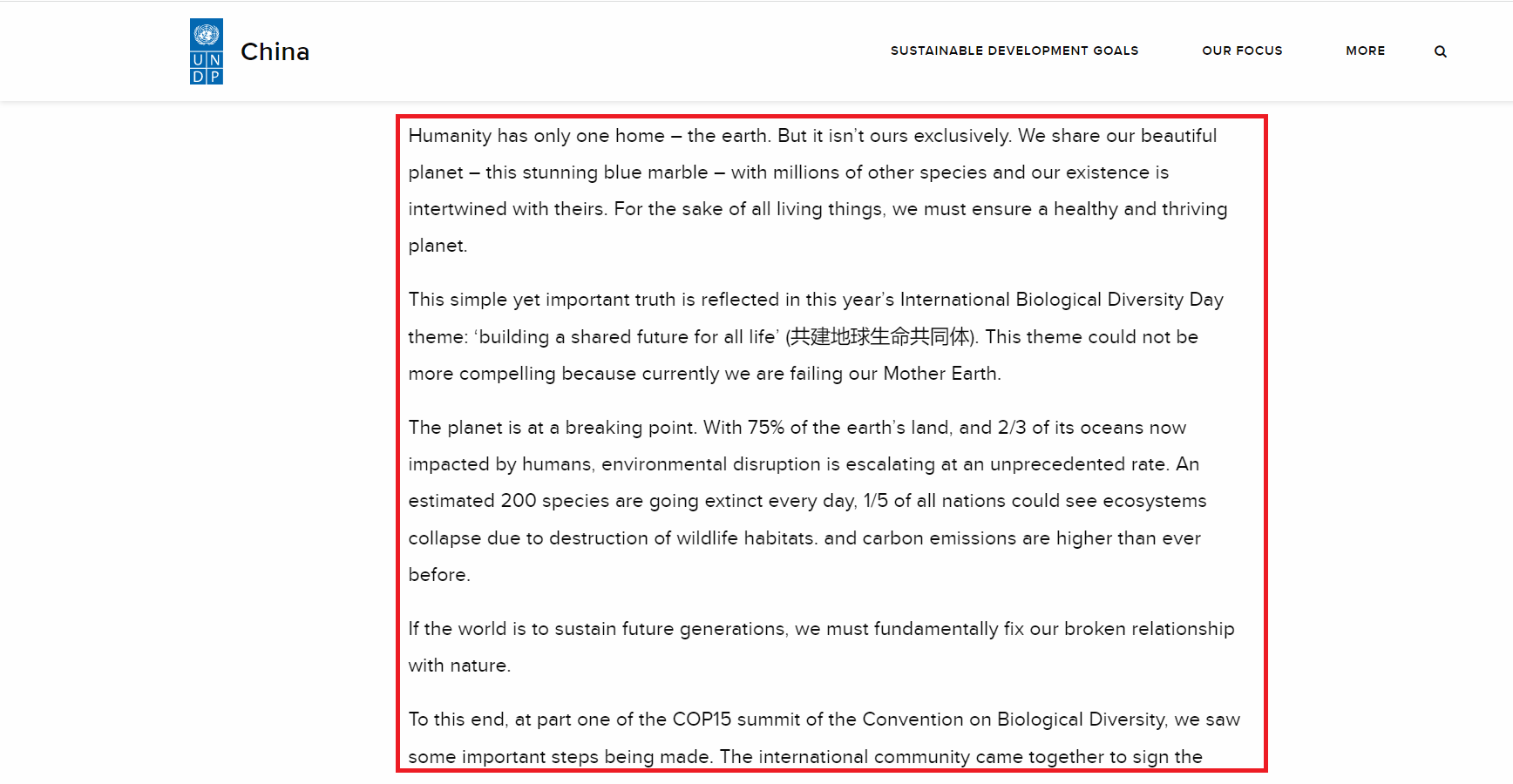
Not all the COs have the same sections. One may have the “press center” URL path inside the website while the other does not. In this case, you need to identify a way to identify which CO (Country Office) has what kind of patterns.

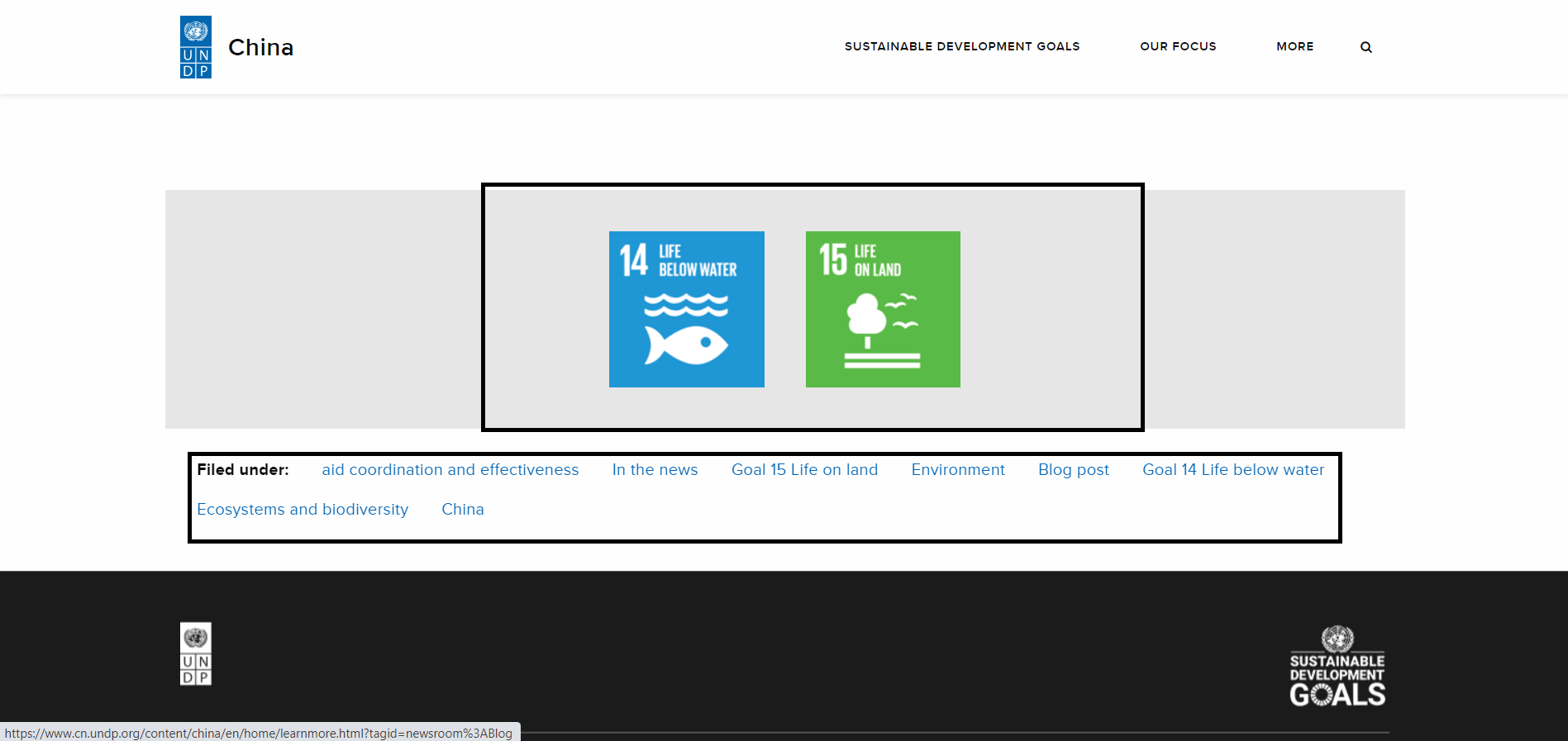
While collecting the data, please collect all the pages without looking for a specific SDG, in this case SDG 16. We would like to scrape all the texts in the blogs, stories etc. To have an SDG labelled text data. At the end of the collection process, we will ask for all the scraped datasets and separately SDG 16 labelled text dataset.

**What are you going to collect over these given URL paths?**

The structure of the posts is similar to each other. You should look for the **“Breadcrumbs”, “Title”, “Postdate” or “Publication date” (which can be in the body of the text), “Text”, “SDG labels”, “Filed Under”.** At the same time do not forget to **specify which CO you scraped the data on, which kind of URL path you scraped the data (blog post, news center etc.), also the URL itself.**







**Challenges:**

* Some of the URL paths have “load more” button which needs to be clicked. If you do not click the load more button you would only scrape a couple of the posts. Since the load more buttons do not affect the URL path, you need to find a programmatical way to click and scrape all the content. To give a preliminary solution, the package” Selenium” does it but you may need to search more about other packages and solutions.
* EXAMPLE LINK: <https://www.gm.undp.org/content/gambia/en/home/blog.html>  
  EXAMPLE LINK(SDG-16): <https://www.gm.undp.org/content/gambia/en/home/blog/2020/undp-support-to-the-establishment-of-the-virtual-courts-in-respo.html>