CPPPC.org

Work Order

# Crawler Description

## *Summary*

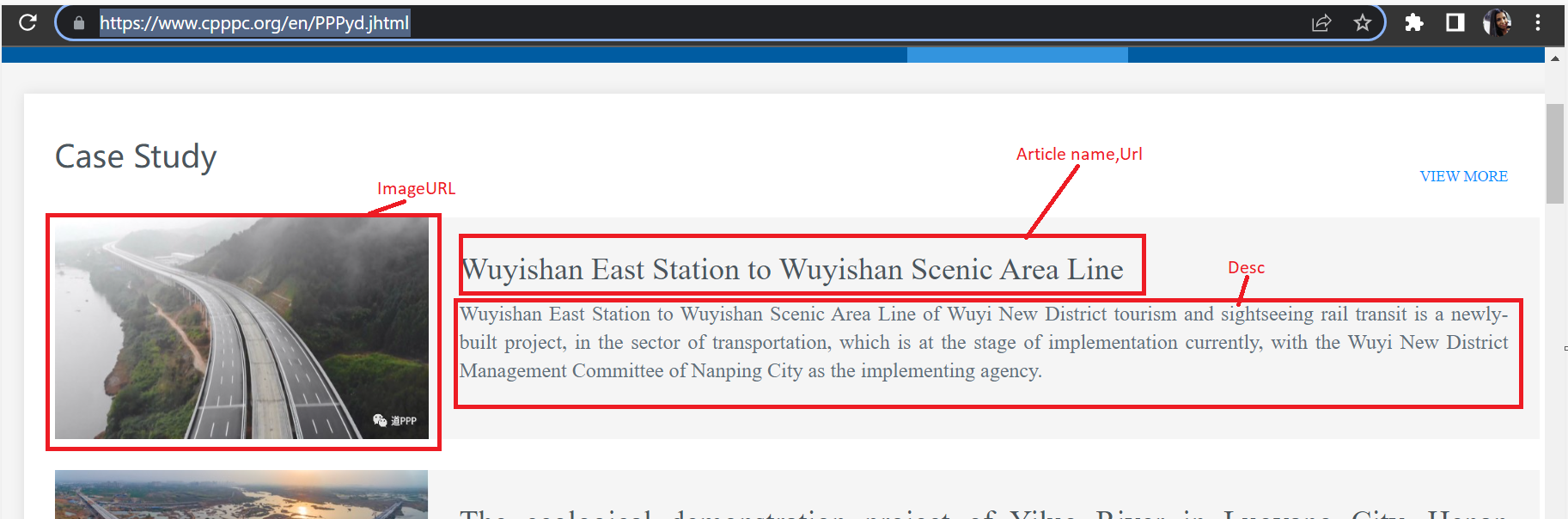
This project will extract the articles which are listed in PPPyd.jhtml portal.  
The goal is to keep the track of article details of all the Lister listed in the PPPyd website.

## *Start URL(s)*

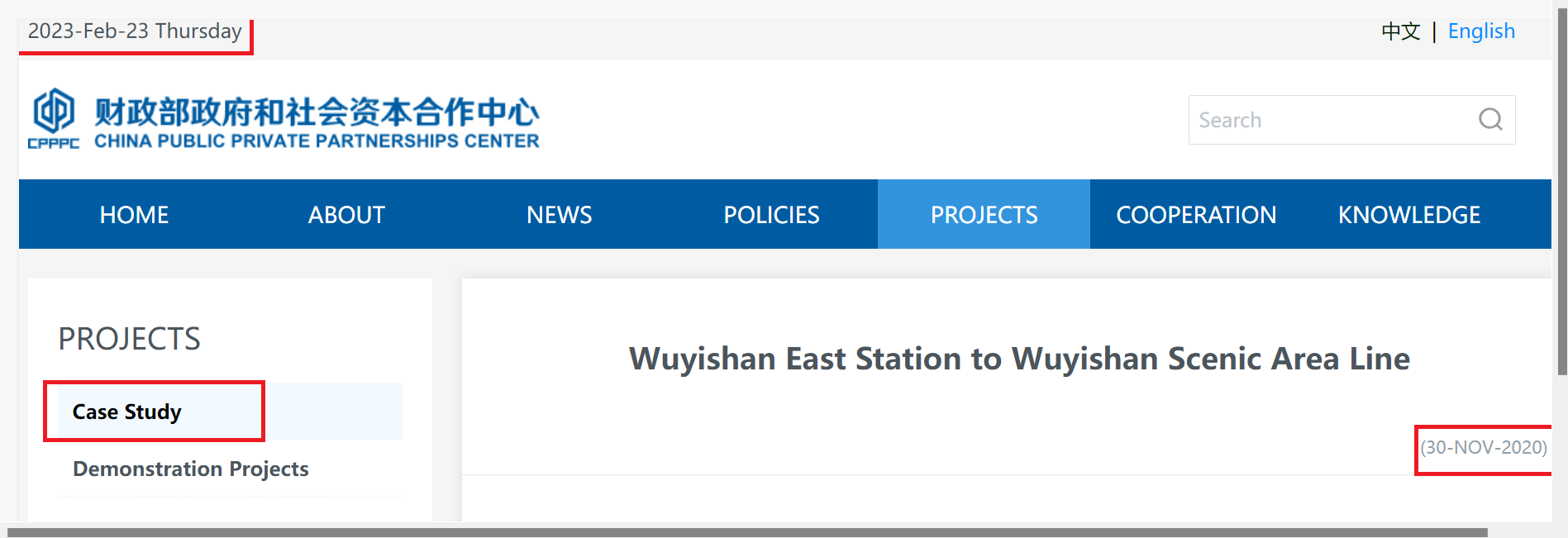
## https://www.cpppc.org/en/PPPyd.jhtml

## *Navigation Requirement(s)*

1. Load up the starting URL ‘http://www.cpppc.org/en/PPPyd.html’
2. Pick details of all the provided articles in listpage page.



1. Iterate through each article to detailpage and pick the information as per requirement.



## *Data Element(s)*

There will be one main crawler that extracts the following fields:  
crawler : task\_cpppc.py

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | Table Schema | |  |  | IMG_256  IMG_257 |
| |  | | --- | | **IMG_258Columns names Sample data** | | | |

|  |  |
| --- | --- |
| Rundatetime | 22-02-2023 18:08 |
| SiteAccessDate | 23-02-2023 |
| ArticleType | Case Study |
| ArticleName | Wuyishan East Station to Wuyishan Scenic Area Line |
| PostDate | (30-NOV-2020) |
| Editor | N/A |
| View | 940 |
| ArticleFrom |  |
| ArticleDesc | Wuyishan East Station to Wuyishan Scenic Area Line of Wuyi New District tourism and sightseeing rail transit is a newly-built project, in the sector of transportation, which is at the stage of implementation currently, with the Wuyi New District Management Committee of Nanping City as the implementing agency. |
| ArticleUrl | [https://www.cpppc.org/en/zlk/999325.jhtml](https://www.cpppc.org/en/zlk/999325.jhtml" \t "https://mail.google.com/mail/u/0/" \l "inbox/_blank) |
| ArticleImageUrl | [https://www.cpppc.org/en/opt/pmo/nfs/images/www/202106/18082726inys.jpg](https://www.cpppc.org/en/opt/pmo/nfs/images/www/202106/18082726inys.jpg" \t "https://mail.google.com/mail/u/0/" \l "inbox/_blank) |

## *Export Format*

The crawler will export one file in CSV format ,Profile Report in HTML format and Matplot Map.  
1. Csv file Format : websitename\_yyyymmdd\_hhmmss.csv

1. Profile Report : websitename\_yyyymmdd\_hhmmss.html
2. Matplot Map: websitename\_yyyymmdd\_hhmmss.png