

**Objective:**

Web site owners want to perform analysis about which hypermedia links are accessed from their web pages and gain information like if found hypermedia links are accessible, served over secure protocol etc. Detailed information that needs to be collected for a web page and each hyperlink found on that web page is stated below. The objective is to provide a solution for this requirement that fulfills the following requirements:

- Developed solution should be provided as a RESTFUL web service for public use
  - User requests the analysis of a web page by providing valid URL to the service as a query parameter
  - Determined/found hypermedia links and information about them should be returned in JSON format
- For a valid URL accessed over HTTP/HTTPS and returning HTML content, an analysis should be performed to collect the following data:
  - Analyzed web page URL
  - Total analysis duration
  - Hypermedia links that are grouped as “internal” and “external”. Internal links are the links whose final URL share the same domain (subdomains should be grouped as external) with the provided web page URL. Others are grouped as “external”.
  - For each link determined by analyzing the document, the following information should be reported:
    - Response code and message
    - Parsed link and final link
    - Available via HTTP or HTTPS (should be decided according to final URL)
    - Result of access attempt
      - If the link is accessible
        - Read contents length
      - Otherwise
        - Information about what went wrong
    - Total access duration (ms)
    - Redirections

**Assumptions:**

For each analysis it should be assumed that remote system will return the same response for the same link. Returned response for a link may differ from analysis to analysis. However, the response will always be the same for the same link in the scope of one analysis.

**Constraints:**

Responsiveness of the service is relevant. Service should not be unresponsive or down, even if remote system fails to respond. Final URL accessed and content should be fetched in 3 seconds. Otherwise, it should be accepted as the URL is not accessible.

For each link maximum allowed redirection count should be defined as 3. If limit is exceeded, it should be accepted as the link is not accessible.

999 should be used as response code for custom errors like invalid content type etc.

**Evaluation:**

Provided solutions will be evaluated according to the following criteria:

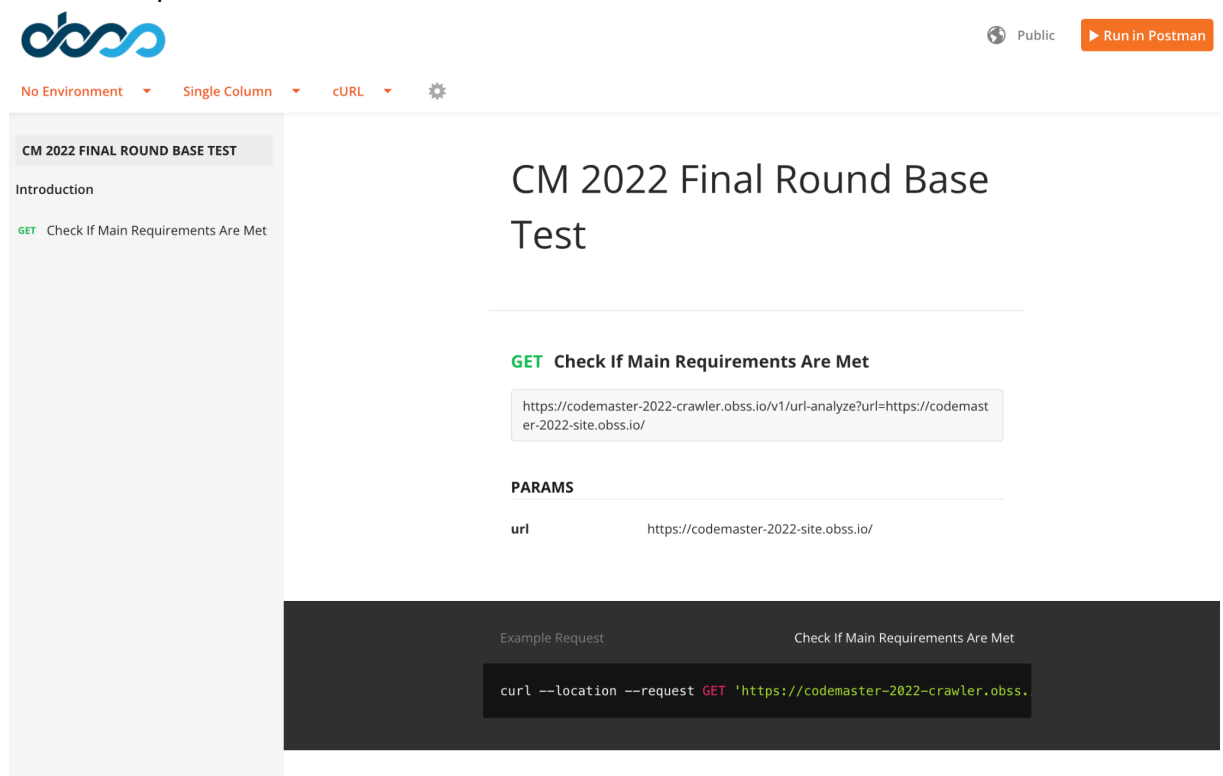
- Correctness
- Performance
- Robustness of the solution
- General code quality and reliability

### Submission:

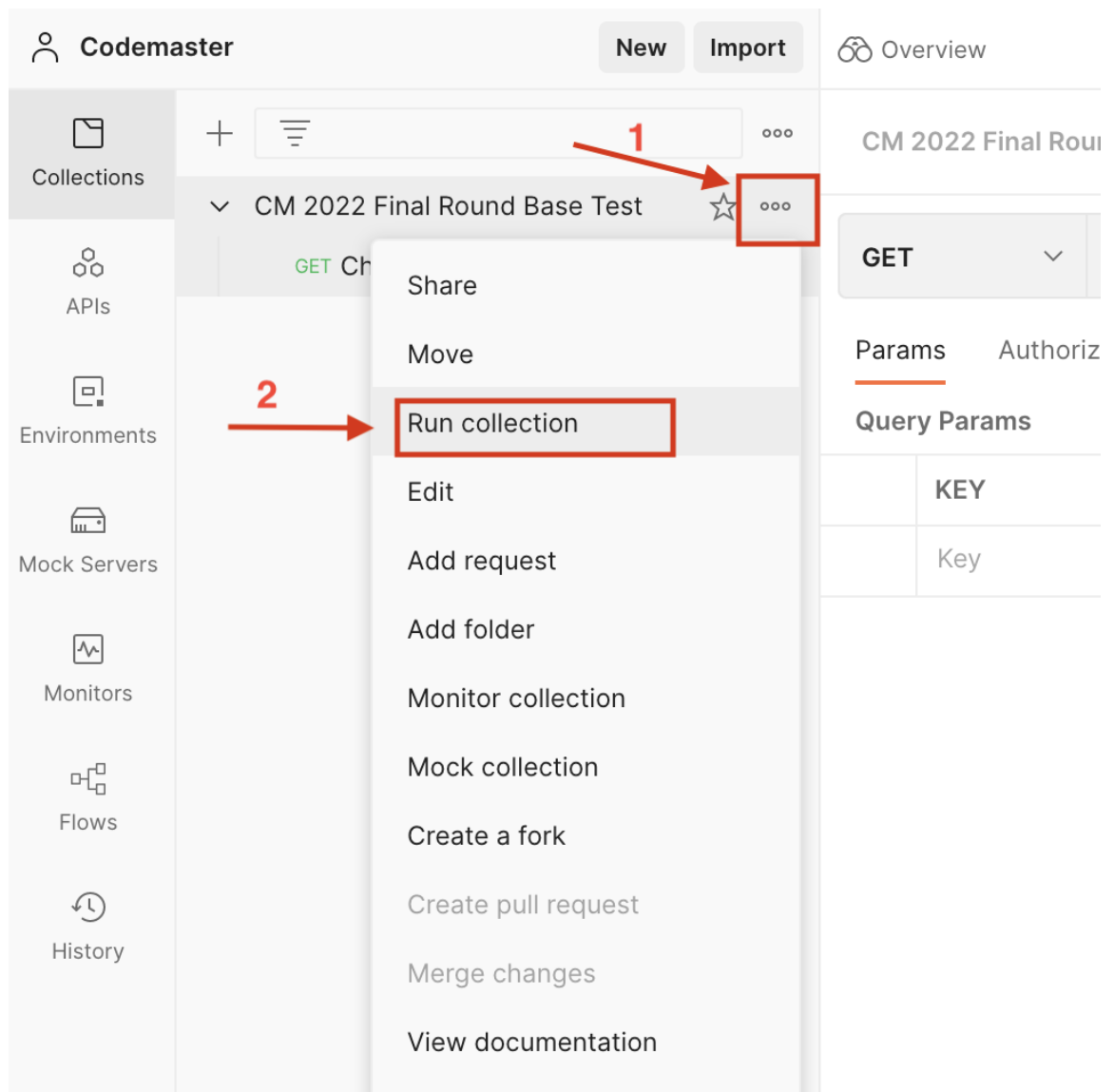
Source code and related documents should be submitted to Git repositories before specified time limit exceeds. Otherwise, the solutions will not be assessed. For each participant, necessary credentials to access Git repository will be provided with an additional document.

### Testing:

- 1) Download Postman for desktop from <https://www.getpostman.com/downloads/>
- 2) Visit <https://documenter.getpostman.com/view/6916431/UVkvHsJX> to download test collection prepared for postman app. Web page similar to the image shown below should be opened



- 3) Click “Run in Postman” button to import test collection to already installed postman application
- 4) Follow below steps to test your solution:
  - a) After clicking the “Play” button, click the “Run” button.



- b) Click the “Run CM 2022...” button from the opened “Collection Runner” window

RUN ORDER Deselect All Select All Reset

☒ **GET** Check If Main Requirements Are Met

Iterations

Delay  ms

Data

☐ Save responses ⓘ

☒ Keep variable values ⓘ

☐ Run collection without using stored cookies

☒ Save cookies after collection run ⓘ

c) After the test is over, you will see the results.

CM 2022 Final Round Base Test CM22\_EXTENDED, just now

View SummaryRun AgainNewExport Results

All TestsPassed (15)Failed (0)

Iteration 1

1

GET Check If Main Requirements Are Methttps://codemaster-2022-crawler.obss.io/v1/url-analyze?url=https://codemaster-2022-site.obss.io/ / Check If Mai...200 OK3296 ms3.136 KB

PassStatus code is 200

PassResponse format is JSON

PassMain content format is OK

PassReached link summary info format is OK

PassInternal link count is OK

PassExternal link count is OK

PassScenario 1

PassScenario 2

PassScenario 3

PassScenario 4

PassScenario 5

PassScenario 6

PassScenario 7

PassScenario 8

PassScenario 9

**Sample Analysis Service:**  
**Request:**

<https://codemaster-2022-crawler.obss.io/v1/url-analyze?url=https://codemaster-2022-site.obss.io/>

**Notes:**

\*\*\*You can test your own web service like:

**localhost:8081?url=**<https://codemaster-2022-site.obss.io/>

\*\*\* Don't forget to change **query param** name if it's different for your service

\*\*\* Don't forget to change **port** with the one you use for your service

**Response:**

```
{
  "url": "https://codemaster-2022-site.obss.io/",
  "analysisDuration": 4109,
  "redirectedURLs": [],
  "responseCode": 200,
  "responseMessage": "OK",
  "internalLinks": [
    {
      "parsedUrl": "http://codemaster-2022-site.obss.io/obss",
      "finalUrl": "http://codemaster-2022-site.obss.io/obssInfo",
      "secured": false,
      "reachable": true,
      "redirectedURLs": [
        "http://codemaster-2022-site.obss.io/obssInfo"
      ],
      "totalAccessDuration": 9,
      "contentLength": 10,
      "responseCode": 200,
      "responseMessage": "OK"
    },
    {
      "parsedUrl": "http://codemaster-2022-site.obss.io/help",
      "finalUrl": "http://codemaster-2022-site.obss.io/help",
      "secured": false,
      "reachable": false,
      "totalAccessDuration": 6,
      "responseCode": 404,
      "responseMessage": "Not Found"
    },
    {
      "parsedUrl": "http://codemaster-2022-site.obss.io/about",
      "finalUrl": "http://codemaster-2022-site.obss.io/about",
      "secured": false,
      "reachable": false,
      "totalAccessDuration": 3006,
      "responseCode": 408,
      "responseMessage": "Request Timeout"
    }
  ]
}
```

```

    },
    {
      "parsedUrl":
"http://codemaster-2022-site.obss.io/post/a-codemaster-codemaster-codema
ster-codemaster",
      "finalUrl":
"http://codemaster-2022-site.obss.io/post/a-codemaster",
      "secured": false,
      "reachable": false,
      "redirectedURLs": [

"http://codemaster-2022-site.obss.io/post/a-codemaster-codemaster-codema
ster",

"http://codemaster-2022-site.obss.io/post/a-codemaster-codemaster",
      "http://codemaster-2022-site.obss.io/post/a-codemaster"
    ],
    "totalAccessDuration": 19,
    "responseCode": 310,
    "responseMessage": "Too Many Redirects"
  },
  {
    "parsedUrl": "http://codemaster-2022-site.obss.io/contact-us",
    "finalUrl": "http://codemaster-2022-site.obss.io/contact-us",
    "secured": false,
    "reachable": false,
    "totalAccessDuration": 10,
    "responseCode": 404,
    "responseMessage": "Not Found"
  }
],
"externalLinks": [
  {
    "parsedUrl": "https://www.google.com",
    "finalUrl": "https://www.google.com",
    "secured": true,
    "reachable": true,
    "totalAccessDuration": 402,
    "responseCode": 200,
    "responseMessage": "OK"
  },
  {
    "parsedUrl":
"http://codemaster-2022-site.obss.io/post/a-codemaster",
    "finalUrl": "https://www.google.com/?gws_rd=ssl",
    "secured": true,
    "reachable": true,
    "redirectedURLs": [
      "http://codemaster-2022-site.obss.io/post/a",

```

```
        "http://www.google.com",
        "https://www.google.com/?gws_rd=ssl"
    ],
    "totalAccessDuration": 646,
    "responseCode": 200,
    "responseMessage": "OK"
},
{
    "parsedUrl": "https://adkakdasd.adasdsad.com",
    "finalUrl": "https://adkakdasd.adasdsad.com",
    "secured": false,
    "reachable": false,
    "totalAccessDuration": 2,
    "responseCode": 503,
    "responseMessage": "Service Unavailable"
},
{
    "parsedUrl": "mailto:olcay.orakci@gmail.com",
    "secured": false,
    "reachable": false,
    "responseCode": 400,
    "responseMessage": "Bad Request"
}
]
```