Web Crawler Project - API Documentation

Author: **Bahy Helmi Hartoyo Putra**
(bahyhelmi97@gmail.com)

Method / Description	**URL**	**Request**	**Response**

[GET]
Crawl Single Website
	http://127.0.0.1:5000/api/v1/crawler?**url**=nyetak.id	-	{
			"broken_link_score": 0.0,
			"contact_us_score": 66.66666666666666,
			"cu_email_exist": 1,
			"cu_phone_number_exist": 1,
			"fraud_score": 0.15846829,
			"link_about_us_exist": 0,
			"link_contact_us_exist": 0,
			"link_tnc_exist": 1,
			"links_response": "{'https://bazar.nyetak.id/index.php?appid=product&cat_id=1': '<Response [200]>', … , ...}",
			//List of link responses
			"merchant_name": "nyetak.id",
			"tnc_refund_policy_exist": 1,
			"tnc_score": 100.0
			}

[GET] Crawl More Than One Website	http://127.0.0.1:5000/api/v1/crawler-batch?**input_file**=datasets/merchants_cleaned.csv&**start_index**=0&**end_index**=10&**output_file**=df_raw	-	"10 line(s) successfully written."

[POST]
Get Prediction Score
	http://127.0.0.1:5000/api/v1/model	{
		"broken_link_score":{
		"0":0.0
		},
		"link_contact_us_exist":{
		"0":1
		},
		"cu_email_exist":{
		"0":1
		},
		"cu_phone_number_exist":{
		"0":1
		},
		"link_about_us_exist":{
		"0":1
		},
		"link_tnc_exist":{
		"0":1
		},
		"tnc_refund_policy_exist":{
			0.15846829

		"0":1 } }	

Notes

- It is not necessary to hit the “**Get Prediction Score**” endpoint since it’s automatically called when hitting the crawling endpoints.
- “**Crawl More Than One Website”** parameter details:
  - **input_file**: link to a .csv file that contains lists of websites to be crawled.
    - example.csv
  - **start_index:** starting index of first website to be crawled in the csv.
  - **end_index:** ending index of last website to be crawled in the csv.
  - **output_file**: name of the outputted csv file, the default saving location is in the *datasets* folder.
    - example_output.csv