Web Crawler Project - Feature List

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

Description

The goal was to create a set of features for the upcoming fraud detection model used in KYB automation process. Before producing these features, discussions have been done with several stakeholders to determine the characteristics of fraud merchants' website observed to this date. Features were extracted by the crawler machine through the crawling process, receiving the website's URL as an input.

Summary

No	Feature	Description	Notes
1	`broken_link_score`	The percentage of link(s) with successful HTTP request (Response 200) out of all sampled hyperlinks. **Values**: [0,100]	• Sampling was done with random sampling with n = 10. • Hyperlinks defined as all links available on the homepage of the website.
2	`link_contact_us_exist`	Flag whether any contact link exists on the website. **Values**: 0/1	-
3	`cu_email_exist`	Flag whether any email address exists on the website. **Values**: 0/1	• Search spaces including homepage and the contact link itself.
4	`cu_phone_number_exist`	Flag whether any phone number exists on the website. **Values**: 0/1	• Search spaces including homepage and the contact link itself.
5	`link_about_us_exist`	Flag whether any about us link exists on the website.	-

		Values: 0/1	
6	`link_tnc_exist`	Flag whether any terms & condition link exists on the website. **Values**: 0/1	-
7	`tnc_refund_policy_exist`	Flag whether any refund policy exists on the website. **Values**: 0/1	• Search spaces including homepage and the terms & condition link itself.

Possible Improvements

- Adding more train data, especially for **REJECTED** merchants. It is good for the model to learn more about the variation of a fraud merchant website.
- Implement content prediction on several pages that considered as content-rich pages. The output of the prediction can be joined together with the current model to enhance the performance of the final prediction, using the ensemble learning method.
- Rediscuss the definition of a fraud merchant and the objective of developing a future model.

Reference

- Source Code
- Beautifulsoup
- Selenium with Python
- Model Validation Notebook
- Hyperparameter Tuning Notebook