Precision - 機能 #15

クロールするドメインの制御

04/16/2019 05:14 AM - 寛正 鴨川

Status:	終了	**Start date:**	04/23/2019
Priority:	通常	**Due date:**	04/23/2019
Assignee:	寛正 鴨川	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour

Description

ページ内のリンクを辿ってクロールする際に、クロール対象のドメイン以外のページを対象から除外する。

バッチ単位での許可するドメインの制御であれば簡単に可能な想定。
起点のURLでの厳密な制御を行うには調査が必要。

対応内容

Scrapy標準で有効になっているOffsiteMiddlewareの機能を利用するために、Spiderのコンストラクタで開始点のURLの取得、許可するドメインの設定を行うようにした。
OffsiteMiddleware:
https://docs.scrapy.org/en/1.6/topics/spider-middleware.html?highlight=offsite#module-scrapy.spidermiddlewares.offsite

起点のURLでの厳密な制御はOffsiteMiddlewareを使う場合に比べて実装量が多くなると判断したため実装していない。

History

#1 - 04/23/2019 04:16 AM - 寛正 鴨川

- Description updated

- Due date set to 04/23/2019

- Status changed from 新規 *to* 終了

- Start date changed from 04/16/2019 to 04/23/2019

- % Done changed from 0 to 100

バッチ単位での制御の実装完了しました。
クローズします。