Precision - 機能 #8

再帰的クローリングの実装

04/08/2019 02:44 AM - 寛正 鴨川

Status:	終了	**Start date:**	04/08/2019
Priority:	通常	**Due date:**	04/09/2019
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URLを抽出し、リンク先を再帰的にクローリングする。

h2. 実装

aタグのhref属性を抽出し、相対パスをURLに変換して新たなリクエストを投げるようにした。
全てのレスポンスに対してこの処理を実行するようにして再帰的なクローリングを実現した。

なお、訪問済みページの重複リクエストの除外、指定ドメイン以外のページへのリクエストの除外はScrapyでカバーされるため実装不要だった。

h2. MT

コーポレートページのトップページ(https://www.interprism.co.jp/) をクローリングしマニュアルテストした。

クローリングでアクセスしたページを出力し、トップページのリンクURLを出力したものと比較したところ、トップページにリンクのない深い階層のページへのアクセスが確認できた。
よって、再帰的クローリングが実現できていると考えられる。
また、外部ドメインのURLはアクセスしていない事を確認した。

History

#1 - 04/08/2019 06:11 AM - 寛正 鴨川

- Status changed from 新規 *to* 進行中

- % Done changed from 0 to 10

着手。

#2 - 04/08/2019 11:10 AM - 寛正 鴨川

- % Done changed from 10 to 50

#3 - 04/09/2019 07:33 AM - 寛正 鴨川

- Description updated

- % Done changed from 50 to 80

#4 - 04/09/2019 07:49 AM - 寛正 鴨川

- Description updated

- Status changed from 進行中 *to* 終了

- % Done changed from 80 to 100