LU13b - Crawler Aufbau Teil 2

WikiPageCollectorService

Der `WikiPageCollectorService` sammelt die URLs von den Seiten, die heruntergeladen werden sollen.

Der Benutzer soll einen Unterordner beziehungsweise einen „Namespace“ angeben können und sämtliche darunterliegenden Seiten sollen rekursiv gesammelt werden.

Leider gibt es dazu keine geeignete und funktionierende XML-RPC-Funktion. Theoretisch könnte man alle Seiten als HTML-Dokumente runterladen und dann sämtliche links (<a href=…) speichern, aber beim DokuWik-Code sind die Links nicht dabei.

In unserem Fall nutzen wir deshalb die Seitenübersicht (https://wiki.bzz.ch/start?do=index). Diese listet sämtliche Namespaces und Seiten hierarchisch auf.

Aber auch hier wird nicht alles auf einmal geladen. Die Inhalte der Ordner wird nur bei Bedarf via `Ajax` dynamisch nachgeladen.

Diese Ajax-Funktion können wir nutzen, um uns die Namespace und Seiten pro Namespace auszugeben.

WikiCrawlerPipelineService

Orchestriert wird das Ganze durch den `WikiCrawlerPipelineService` . Ebenfalls können hier ergänzende Funktionen wie Filterungen eingebaut werden.

Zum Beispiel müssen bereits gespeicherte Seiten nicht erneut heruntergeladen werden, wenn man davon ausgeht, dass sich die Seite in der Zwischenzeit nicht geändert hat.

Ebenfalls kann man Seiten, die lediglich zur Navigation dienen, theoretisch ignorieren, da diese Seite kein nützliches Wissen enthalten.

Beispiel: https://wiki.bzz.ch/de/modul/ffit/3-jahr/java/learningunits/lu01/start