Семинарское задание №1
Обойти сайт

Мосиенко Константин Викторович

2019

Необходимо написать простой робот и обойти часть сайта

`https://simple.wikipedia.org/`

Конечной целью является коллекция статей без лишних страниц (без картинок, обсуждений, каталогов, страниц редактирования и т.д.), поэтому стоит подумать о фильтрации и нормализации урлов во время обхода. Подумайте о политике вежливости и ограничьте количество запросов в секунду (начните с одного запроса в секунду), в противном случае вас забанят. В качестве сида возьмите `https://simple.wikipedia.org/wiki/Main_Page`.

Сохраните полученную коллекцию в виде набора HTML документов - она пригодится в других заданиях. Для хранения статей необходимо использовать класс FileStorage, данный в задании. Он представляет собой key-value хранилище, располагающееся на диске.

На сайте очень много документов, которые даже не стоит скачивать. Чтобы их обнаружить, пишите логи и ищите там урлы, не похожие на статьи (шаблоны, разговоры и т.д.). Но есть одно исключение - страницы категорий. Их обязательно надо качать, так как они являются отличным источником ссылок на другие статьи.

Сравните размер вашей коллекции с показателем на главной странице сайта (на момент написания - 142749). Сколько статей вы не нашли? Сколько нашли лишних?

Не затягивайте с началом обхода - если вас забанят в последний момент, будет обидно. Мне присылайте *.ipynb с вашим кодом, а также *.dict файл получившейся коллекции.