Семинарское задание №6
Page Rank

Мосиенко Константин Викторович

2017

Необходимо написать простой робот и обойти сайт

`http://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. Фильтрацию удобно делать с помощью регулярных выражений. 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 `http://simple.wikipedia.org/wiki/Main_Page`.

Сохраните полученную коллекцию в виде набора HTML документов - она пригодится в других заданиях. Подумайте над периодическим сохранением состояния робота на диск, чтобы его можно было перезапустить в случае падения. Как минимум стоит сохранить множество обойдённых и множество найденных урлов.

Сравните размер вашей коллекции с показателем на главной странице сайта (на момент написания - 123627). Сколько статей вы не нашли?

По полученной коллекции вычислите *ранг господина Пейджа* для каждой страницы и приведите топ-20 страниц.

Составьте отчёт с описанием процедуры фильтрации, нормализации и обхода. Укажите размер коллекции и приведите топ. Пришлите мне код и отчёт.

Не затягивайте с началом обхода - если вас забанят в последний момент, будет обидно.