Информационный поиск
Домашнее задание 1

March 3, 2016

В рамках выполнения первого домашнего задания вам предлагается написать собственный поисковый робот (Web Crawler). Для решения сложных технических задач (скачивания документа по урлу, парсинг разметки) разрешается использовать сторонние библиотеки. Ваш поисковый робот должен скачать сайт `http://simple.wikipedia.org/wiki/` без перехода по внешним ссылкам. Также нас будут интересовать только статьи с этого сайта (картинки, комментарии и др. должны быть проигнорированы). Скачанные страницы необходимо сохранить в отдельной дирректории docs/ следующим образом: файлы с контентом имеют имена 1.html, 2.html, 3.html и т.д., а в дополнительном файле urls сохранено отображение $n.html \to url_n$. По одной разделённой символом табуляции паре на строку. Воспользовавшись парсером HTML, необходимо для каждого документа n.html получить файл n.txt, содержащий только текст (отсутствует разметка). Если есть реализация поискового индекса с семинарского занятия, запустите его на скачанной коллекции.

Помимо получения коллекции документов необходимо вычислить следующие характеристики:

- построить гистограмму распределения размеров текстовых документов в байтах;
- построить гистограмму распределения in/out степеней вершин ссылочного графа;
- для каждого документа вычислить его расстояние от главной страницы в кликах, построить гистограмму распределения расстояний;
- для каждого слова вычислить частоту его появления в коллекции и построить гистограмму распределения частот (видимо, придётся использовать логарифмические оси);
- на основе ссылочного графа посчитать PageRank документов и привести топ-20 результатов;

Полученные результаты нежно сформулировать в отчёте с соответствующими пояснениями и выводами, которые можно сделать на основе полученных данных. Мне(я семинарист) на почту присылайте архив с исходным кодом программ, файл urls, и отчёт. Обязательно следуйте рекомендациям по оформлению решений со страницы курса.