Домашнее задание №1

Виталий Емельянов

30 марта 2016 г.

Вычислительные возможности

В качестве языка программирования использовался Python 2.7, а также сторонние библиотеки: requests, beautifulsoup, numpy и matplotlib.

Все вычисления проводились на ноутбуке с 4 Гб ОЗУ и 3-х ядерным процессором AMD Phenom II P860.

Вычисления

Поисковый робот

- Реализован в классе WebCrawler в файле web_crawler.py
- Обход сайта simple.wikipedia.org начинается с 7 страниц и проводится, соответственно, в 7 потоков с общей очередью и двумя общими словарями: словарем посещенных страниц и словарем графа в представлении списков смежности
- Поиск ссылок проводился только в блоке div класса mw-body-content, так как именно эта часть html кода страницы, судя по всему, отвечает содержанию статьи. Учитывались только внутренние ссылки, то есть ссылки, начинающиеся на /wiki/ без символа : после, так как именно такого вида URL имеют страницы категорий

  Все, что в URL страницы находилось после символа # игнорировалось
- Количество статей, которое удалось обойти равняется 126738
- Граф сайта сохранялся в виде списков смежности. Допускались петли, но не кратные ребра

- По времени обход сайта занял около 140 минут. Было задействовано около 2 Гб ОЗУ и почти все процессорное время

Обработка результатов

- Реализована в файле wikipedia_parser.py и report.py
- Парсинг html в текст статьи проводился над содержимым блока div с id = mw-content-text
- Словом считалась любая последовательность букв или цифр
- PageRank вычислялся с damping factor=0.85 и количеством итераций равным 100

Результаты

Построим гистограммы, обозначив медианы распределений

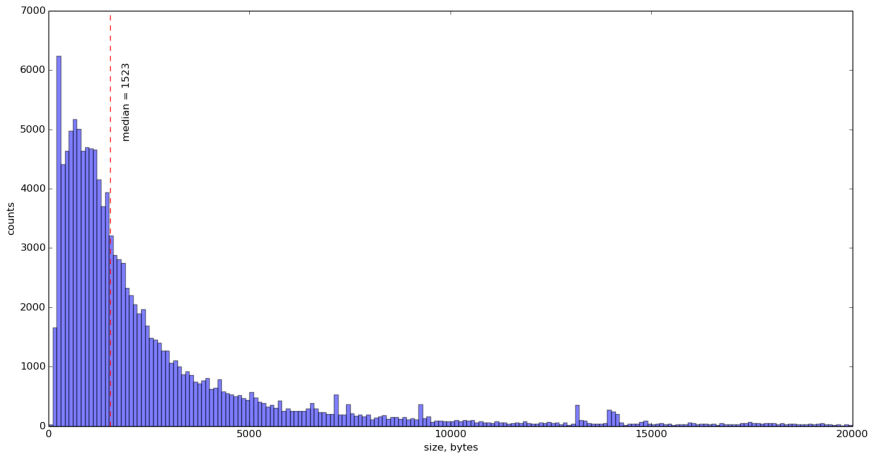

Рис. 1: Гистограмма распределения размеров текстовых документов в байтах

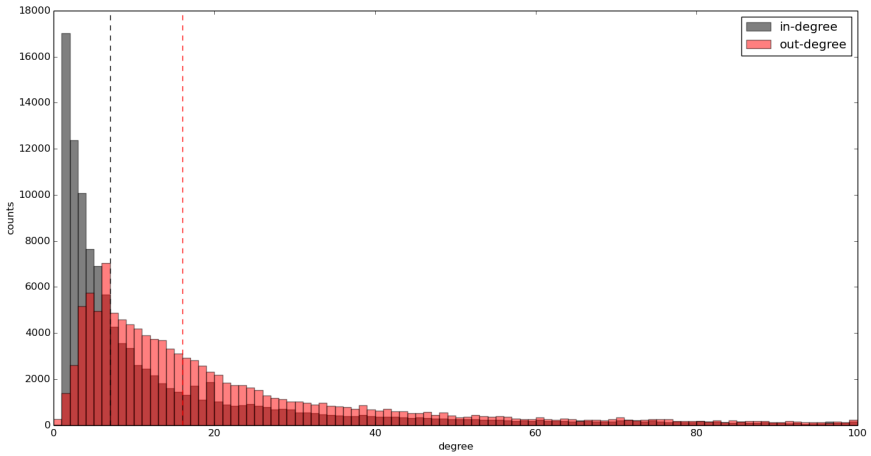

Рис. 2: Гистограмма распределения in/out степеней вершин ссылочного графа

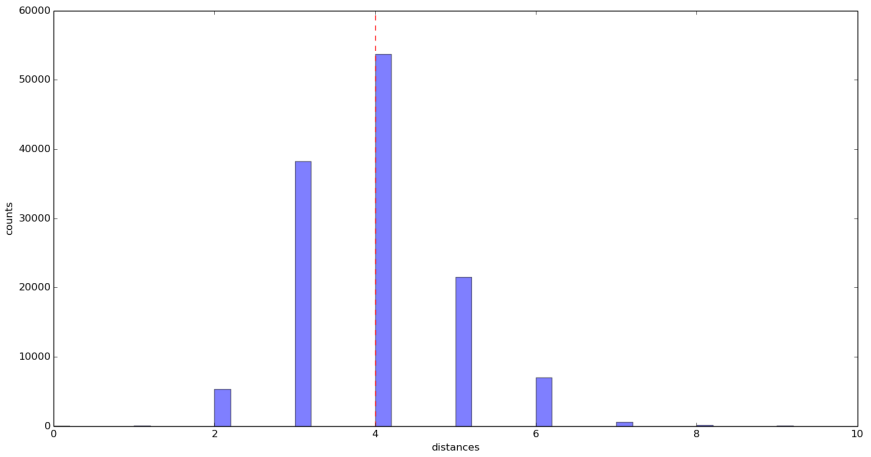

Рис. 3: Гистограмма распределения расстояний от главной страницы

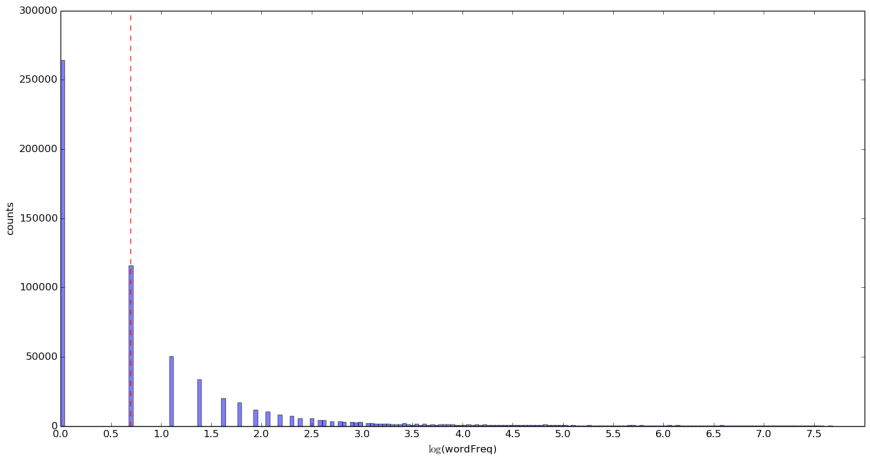

Рис. 4: Гистограмма распределения частот слов в коллекции

Таблица 1: Топ-20 страниц с наибольшим PageRank

simple.wikipedia.org/wiki/United_States	0.00350605239562
simple.wikipedia.org/wiki/Multimedia	0.0034723408902
simple.wikipedia.org/wiki/France	0.00205203712935
simple.wikipedia.org/wiki/United_Kingdom	0.00128681867597
simple.wikipedia.org/wiki/English_language	0.00121120081826
simple.wikipedia.org/wiki/City	0.00112932415892
simple.wikipedia.org/wiki/England	0.00107643589745
simple.wikipedia.org/wiki/International_Standard_Book_Number	0.0010663529374
simple.wikipedia.org/wiki/Geographic_coordinate_system	0.00104752916493
simple.wikipedia.org/wiki/Japan	0.0010475247411
simple.wikipedia.org/wiki/Association_football	0.00103185693803
simple.wikipedia.org/wiki/Country	0.00102592069692
simple.wikipedia.org/wiki/Wikimedia_Commons	0.000935381107742
simple.wikipedia.org/wiki/Americans	0.000925460339453
simple.wikipedia.org/wiki/Europe	0.000908909209993
simple.wikipedia.org/wiki/Definition	0.000905201913148
simple.wikipedia.org/wiki/Canada	0.000898056499379
simple.wikipedia.org/wiki/Departments_of_France	0.000888681594356
simple.wikipedia.org/wiki/Germany	0.000870238162516
simple.wikipedia.org/wiki/Australia	0.000825070333764

Из таблицы видно, что Multimedia имеет высокий PageRank. На самом же деле это происходит потому, что большая часть страниц имеет плашку Wikimedia с ссылкой на статью simple.wikipedia.org/wiki/Multimedia

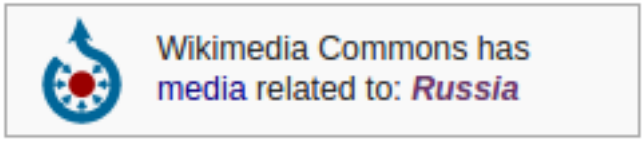

Рис. 5: Плашка Wikimedia

То есть, попадание данной страницы в топ-20 в какой-то мере ошибочно и связано с тем, какой мы критерий выбираем для отбора ссылок.

Похожая ситуация происходит и со статьей про ISBN. Действительно, в статьях, посвященных книгам, указывается номер ISBN книги и ссылка на то, что же такое ISBN.

Под подозрение попадает также и статьи про Wikipedia Commons и Geographic coordinate system. Очевидно, что в топ-20 они попали по примерно той же причине, что и предыдущие два примера.

Про остальные статьи из списка можно сказать, что это довольно естественно, что они туда попали.