Домашнее Задание №1

Web Crawler

ученицы 2 курса

отделения Компьютерных Наук

Тихоновой Марии

В задании нам требовалось написать собственный поисковый робот Web Crawler, который бы скачал статьи с сайта *http:// simple.wikipedia.org/wiki/* без перехода по внешним ссылка.

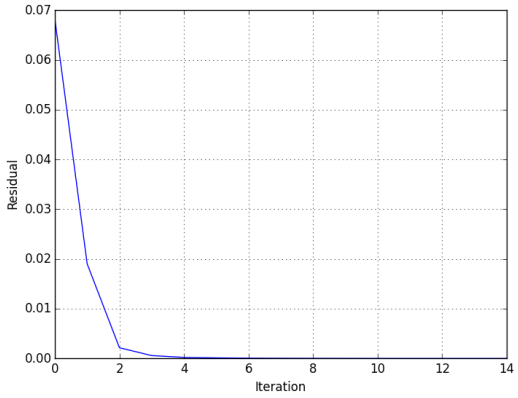
Я написала поискового робота (документ *web_crawler.py*), который, начиная с главной страницы сайта переходит по ссылкам, ведущим на статьи, скачивает содержимое страниц (пока без обработки и извлечения текста), сохраняет содержимое статей в файл *n.html* (*n* – номер прочтенного файла), записывает отображение $n.html \rightarrow url_n$ в файл *urls.map*, а также для каждой посещенной страницы записывает список статей, на которые она ссылается в файл *links.map*. ($page\_url \rightarrow url_1, url_2, url_3$ …).

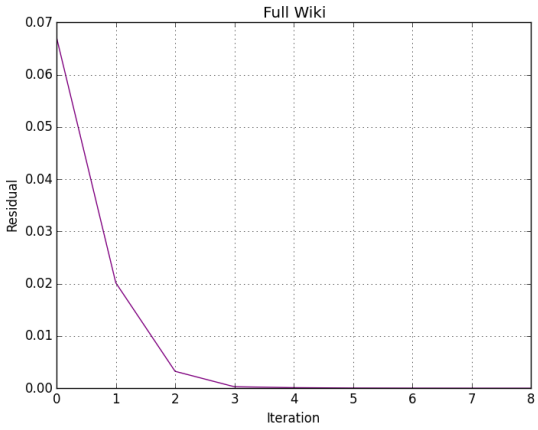
Извлечение текстов статей выполняется файлом *getting_text.py*. Программа для каждого скаченного файла $n.html$ извлекает из него текст статьи и записывает его в файл $n.txt$.

Теперь пришло время посчитать PageRank для каждой страницы, построить различные графики и сделать выводы. Здесь у меня возникли технические сложности, связанные с нестабильной работой интернета. Мне очень стыдно, но в последнее время у меня в доме интернет постоянно прерывается. Поэтому скачать *simple.wikipedia* целиком у меня не получилось. После нескольких неудачных попыток, нескольких суток перезапусков, я скачала около 9000 статей (это связано не с тем, что моя программа работает некорректно, а с тем, что у меня может неожиданно отключится интернет). Однако выполнить задание мне все равно было очень интересно! Поэтому я решила сделать некое мини исследование, провести сравнительный анализ: насколько данная выборка репрезентативна и насколько данное частичное исследование сайта позволяет точно построить PageRank. Я попросила ребят, у которых получилось выкачать все статьи, перекинуть мне свою коллекцию файлов *n.html* и сравнила результаты работы моей программы, полученные на полной коллекции документов, с результатами работы программы, полученными на моей, усеченной, коллекции. Результаты получились очень интересными.

Обработка усеченной коллекции статей производится файлом *irhw1_small.py*, а полной - *irhw1_full.py*. Гистограммы, относящиеся к усеченной коллекции нарисованы синим, а к полной – фиолетовым.

Самым долгим при работе программы был подсчет расстояния от главной страницы, однако после того как граф построен и глубина посчитана, PageRank сходится достаточно быстро (изобразила сходимость на графиках):





.

На маленькой коллекции получилось больше итераций поскольку там я поставила большую точность ($Eps = 10^{-6}$), а на полной $Eps = 10^{-5}$.

Теперь посмотрим, какие сайты вошли в топ 20. (Привожу названия статей и их PageRank.

ТОП 20 (small collection):

```
Article:  Main_Page        PageRank:  0.04727538304221646
Article:  Multimedia       PageRank:  0.008385104008015915
Article:  United_States    PageRank:  0.0062505098761633
Article:  Country          PageRank:  0.002879076704288289
Article:  English_language         PageRank:  0.0027543872694095195
Article:  Movie   PageRank:  0.0027408676330385497
Article:  Definition       PageRank:  0.002717958179583026
Article:  France           PageRank:  0.002535667623260282
Article:  International_Standard_Book_Number        PageRank:  0.002357234243165733
Article:  United_Kingdom   PageRank:  0.002346628648671306
Article:  Canada           PageRank:  0.002242004559004959
Article:  Government       PageRank:  0.002169018420360522
Article:  Television       PageRank:  0.0021440593625877366
Article:  Internet_Movie_Database          PageRank:  0.002141613605420426
Article:  Europe           PageRank:  0.002046271940764577
Article:  Mathematics      PageRank:  0.0019287313183231
Article:  Language         PageRank:  0.0017499987122459237
Article:  Law     PageRank:  0.0016485080942612936
Article:  Germany          PageRank:  0.0016459977308724782
Article:  Americans        PageRank:  0.0016372470631995054
```

ТОП 20 (full collection):

```
Article:  Main_Page        PageRank:  0.043315536211216095
Article:  Multimedia       PageRank:  0.005655534082723838
Article:  United_States    PageRank:  0.003532264545945165
Article:  Definition       PageRank:  0.0024626391591453215
Article:  France           PageRank:  0.0021651695094869936
Article:  English_language         PageRank:  0.0018705865562235878
Article:  England          PageRank:  0.0017721933825950408
Article:  United_Kingdom   PageRank:  0.0017102925362380479
Article:  Europe           PageRank:  0.0016917626199318707
Article:  International_Standard_Book_Number        PageRank:  0.0016430650389202227
Article:  Country          PageRank:  0.0016227152566438499
Article:  Government       PageRank:  0.0013869286335118753
Article:  Germany          PageRank:  0.0013081045212065534
Article:  Japan   PageRank:  0.0012714189027780193
Article:  Movie   PageRank:  0.0012694285792197555
Article:  Television       PageRank:  0.0012342851067777763
Article:  Animal           PageRank:  0.0012299885594538428
Article:  Wikimedia_Commons        PageRank:  0.0012019535088415986
Article:  Mathematics      PageRank:  0.0012011574313878664
Article:  Spain      PageRank:  0.0011963272614410798
```

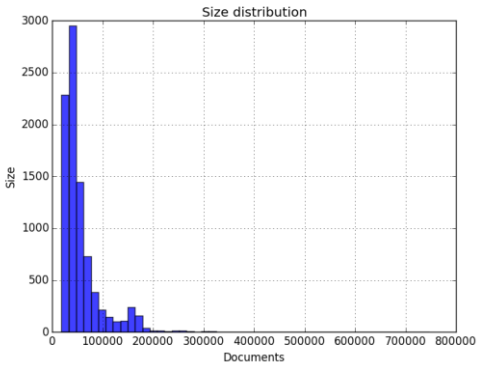
Оказалось, что даже неполный набор сайтов позволяет достаточно адекватно построить PageRank для самых популярных страниц. Топ-3 полностью совпадает, дальше начинаются небольшие расхождение, однако они заключаются больше в порядке следования статей, а списки самих статей примерно совпадают. Интересно заметить, что в силу устройства робота даже при неполном исследовании сайта самые популярные статьи оказываются скачанными и обработанными, а их PageRank отражается адекватно. Это очень хорошо!

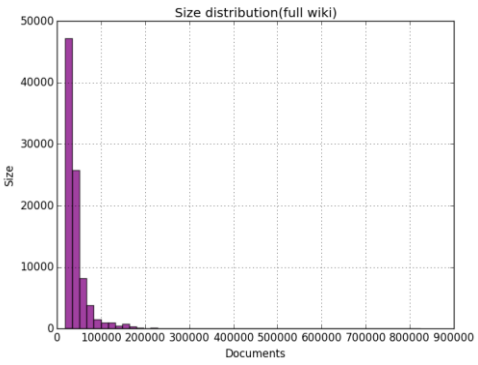
Что можно сказать про сам Rank. Понятно, что я большим отрывом лидирует главная страница, за ней следуют *Multimedia* и *United States*.

Интересно, что в обоих топах присутствует много статей, посвященных различным странам: *United States, France, England, United Kingdom, Germany, Japan, Spain.*

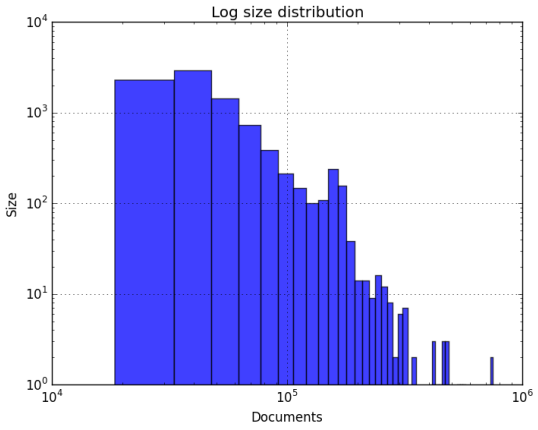
Построение графиков.

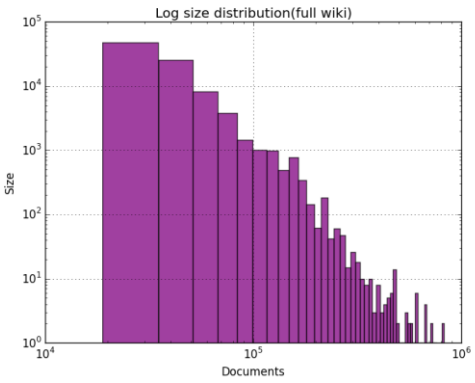
Построим гистограмму распределения размеров текстовых документов в байтах.





и в логарифмической шкале:





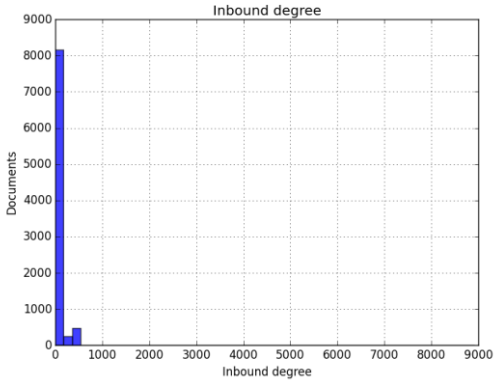
как мы видим, распределения примерно совпадают.

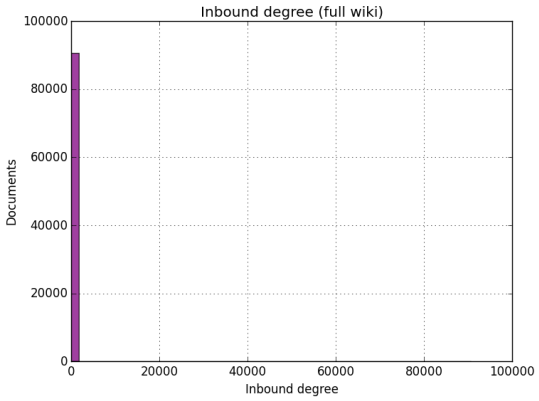
Посмотрим на самые длинные статьи.

По большой коллекции это `https://simple.wikipedia.org/wiki/List_of_Detroit_Red_Wings_players` - `825060 байт.`

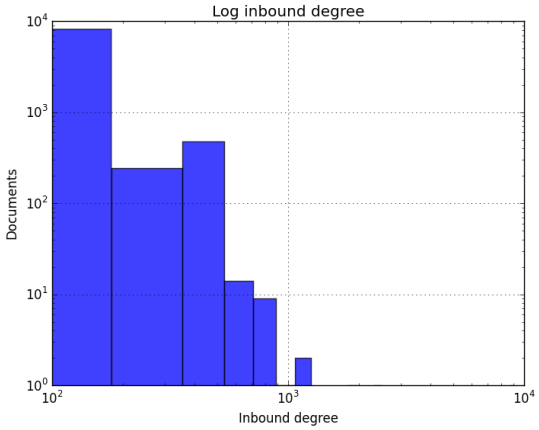
По маленькой `http://simple.wikipedia.org/wiki/2012_in_film` - `747260 байт.`

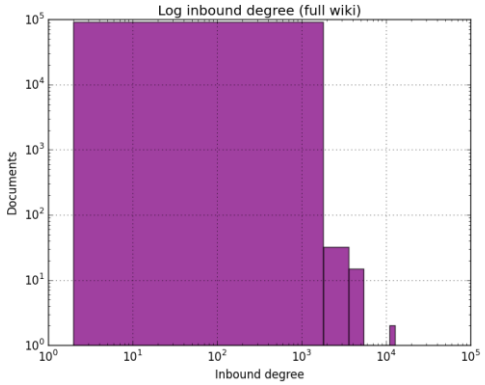
Построим распределение in/out степеней вершин ссылочного графа.





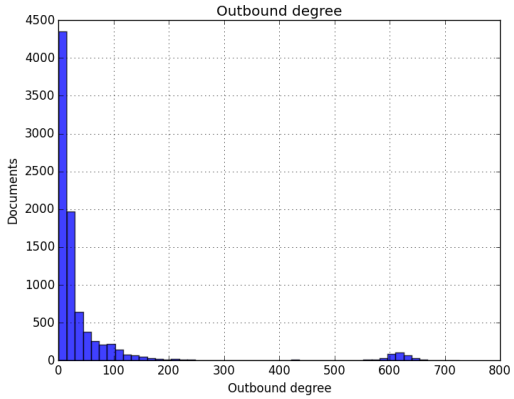
Ничего не понятно, посмотрим на логарифмическую шкалу:

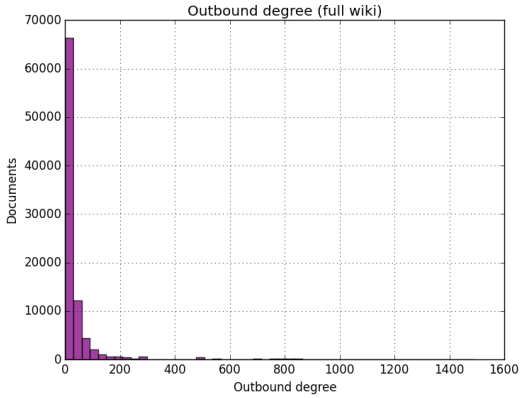




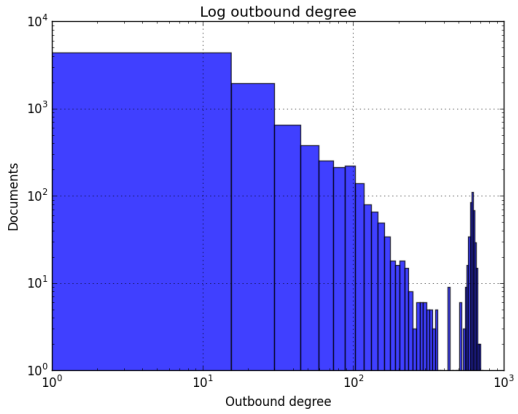
.

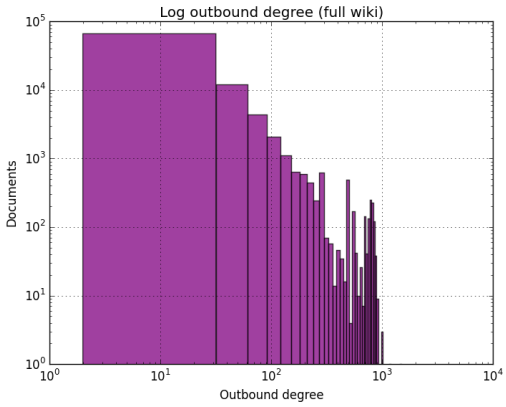
Здесь графики несколько различаются, что понятно, на неполной Вики у нас число ребер меньше, поэтому и графики отличаются. Интересно, что на главную страницу (Main Page) ссылаются абсолютно все, включая ее саму.





В логарифмической шкале:



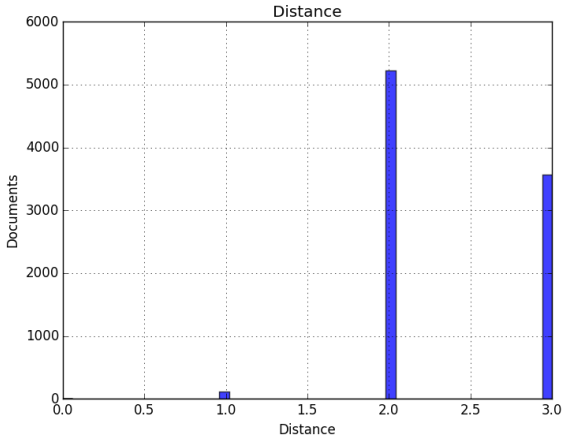


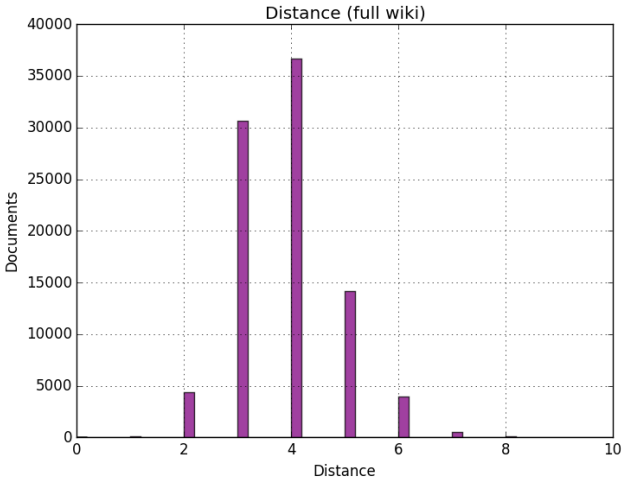
.

Интересно, что рекордсменом по числу ссылок является статья: https://simple.wikipedia.org/wiki/List_of_years.

Количество ссылок на ней просто зашкаливает: 1489. В статье содержится список годов со ссылками на соответствующие статьи.

Построим распределение от главной страницы в кликах:

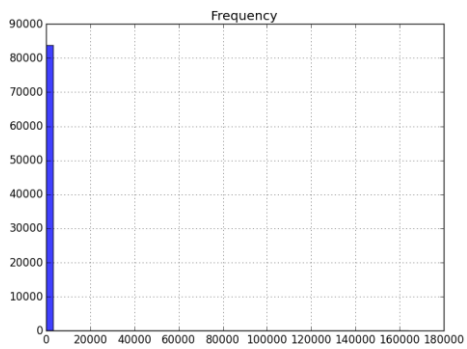


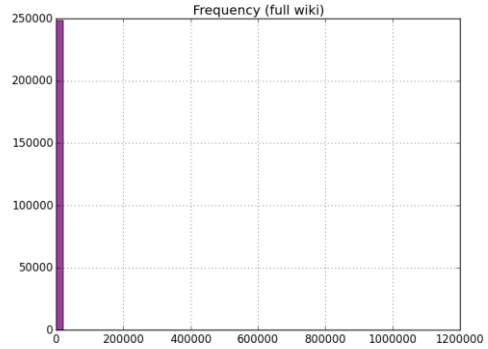


Здесь моя выборка уже не репрезентативна: поскольку у нас осуществлялся обход в ширину, то сначала посещались самые близкие статьи, а до дальних

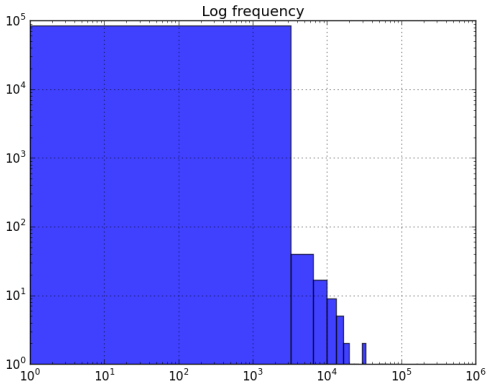
робот добраться не успел. По большой выборке видно, что в среднем статьи находятся на расстоянии 3-4 кликов от главной страницы и в принципе за 10 кликов мы сможем попасть даже на самую дальнюю статью.

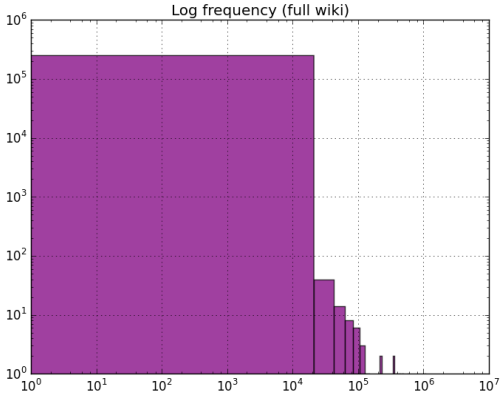
Частота появления каждого слова в коллекции:





Эм… Ничего не понятно, попробуем в логарифмической шкале:





Так лучше.

Выводы.

По полученным на двух выборках данным можно сказать, что даже при неполном исследовании сайта поисковый робот выдает неплохие результаты. Причем, благодаря тому, что он использует поиск в ширину, то он в первую очередь посещает самые популярные страницы, на которые больше всего ссылок. Это хорошо. Шанс того, что пользователь будет запрашивать запрос именно по этим статьям большой. Но, конечно, это меня не оправдывает. Скачивать надо все целиком.

Понятно, что по популярности лидирует Main Page. Интересно, что в топе присутствует много статей, посвященных различным странам: United States, France, England, Germany и т. д.

Также заметим, что большинство статей находится на расстоянии меньше 10 кликов от главной страницы.