Информационный поиск

Домашнее задание 1

Отчет “WebCrawler”

В данном задании требовалось написать скрипт-паука, который бы выкачал сайт

http://simple.wikipedia.org/wiki. И выполнил сопутствующие операции и подготовил данные для анализа.

Основной скрипт был написан на языке Python. Для работы с ссылками и загрузки html-страниц использовалась библиотека urllib2(а так же, для анализа ссылки и помощи в дальнейшем анализе, библиотеки: urlparse, re, SoupBeautiful,…)

Для анализа полученного html-документа(в частности выделения осмысленного текста без тегов), а так же для удобного получения списка всех ссылок, встречающихся в данном html-документе использовалась библиотека SoupBeautiful.

Описание основных деталей алгоритма.

- Алгоритм основан на поиске в ширину по сетевому графу, где страницы – вершины, а ссылки – переходы между ними.
- Для каждой рассматриваемой страницы, и для каждой ссылки ведущей с него ссылка проверяется следующим образом: дополняется до полного пути(если есть необходимость), далее отсекается все что идет после символа ‘#’(то есть убираем указатель на место на странице) Выделяем host и сравниваем с simple.wikipedia.org, а так же проверяем наличие каталога /wiki. Далее проверяем на наличие специфических расширений (как .jpg, .png…). А так же убираем ссылки ведущие в каталоги “…/Catalog:…”. Кроме этого, замечаем что страницы на редактирование, комментарии и т.д содержат символы **: , . , ?** Проверяем путь ссылки на эти символы. В противном случае ссылку можно считать валидной, и исходя из наблюдений, все такие ссылки действительно ведут на какую-то статью википедии.
- После этого проверяется, не была ли посещена страница ранее.

Ввиду большого количества страниц, скачиваемых алгоритмом. В каждом файле “n.html”/”n.txt” мы храним текст полученный сразу с нескольких ссылок(от 10 до 25). Соответственно все данные приведены в файле “URLS.TXT”

- Во время работы алгоритма также, естественным образом, поддерживаются массивы расстояний от вершины(исходя из свойств поиска в ширину – полученное расстояние до страницы является найменьшим возможным количеством клик от начальной страницы, что бы ее достичь).
- Массивы входная и выходная степень вершины. А так же некоторые вспомогательные массивы.
- Явное задание графа в виде списков смежности.

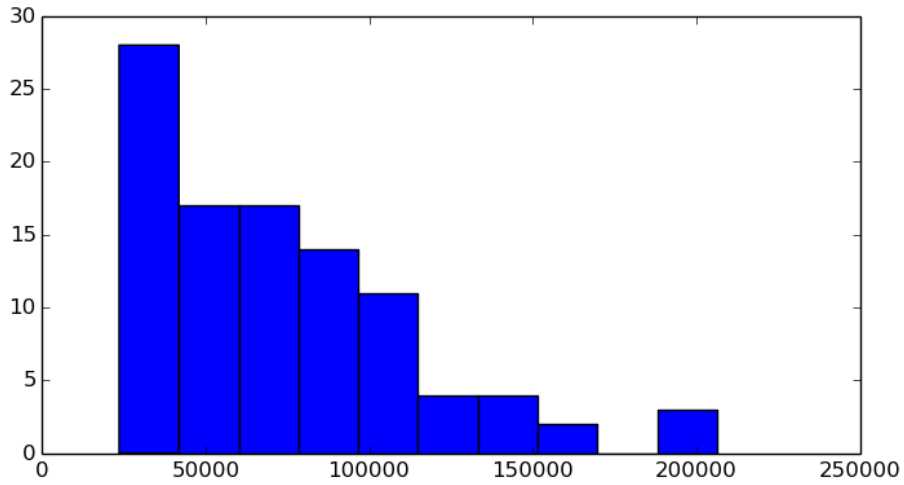
- Все собранный данные записываются в файлы. Все файлы приведены в архиве вместе с этим отчетом.
- Для подсчета PageRank-а страниц после выполнения основного алгоритма строится матрица перехода в явном виде (для N = 1000). Сами же значения PageRank вычисляются итеративно.
- Реализация поискового индекса (даже с TF-IDF), а так же подсчет частоты вхождения для каждого слова реализован на C++. Все исходные коды и сопутствующие файлы так же приложены.

НО, на основании слишком большого размера(и, что немаловажно, длительной работы) количество скачиваемых страниц приходится ограничивать вручную.

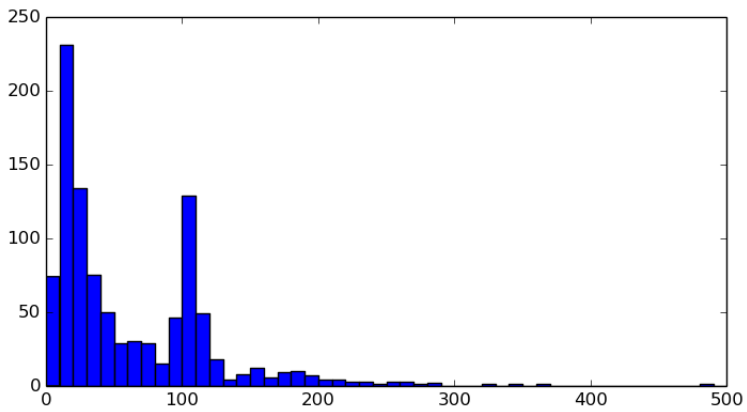
Рассмотрим теперь несколько вариантов:

СКАЧИВАЮТСЯ ПЕРВЫЕ 1000 СТРАНИЦ ВИКИПЕДИИ.

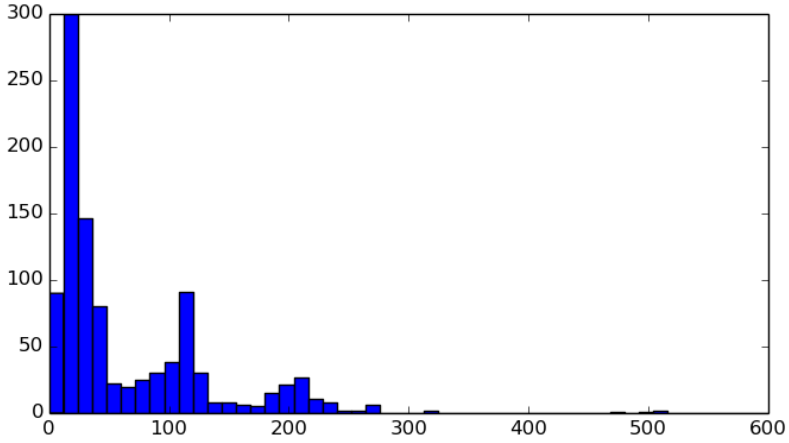
- **Количество рассматриваемых алгоритмов страниц ограничено числов 1000. Данные скачаны 13.04. Остальные страницы не рассматриваются. Но аккуратно обрабатываются все внутренние ссылки и прочее.**
- **В каждом файле “n.txt” сохраняется по 10 страниц. Средний вес одного файла 50KB. Время работы алгоритма до 5 мин.**
- **Гистограмма распределения размера текстовых документов(байт):**



- **Гистограмма распределения in/out степеней вершин ссылочного графа:**

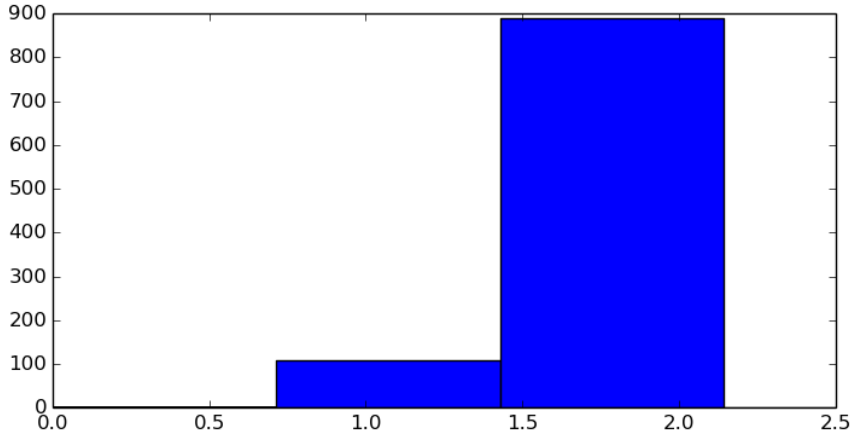


IN

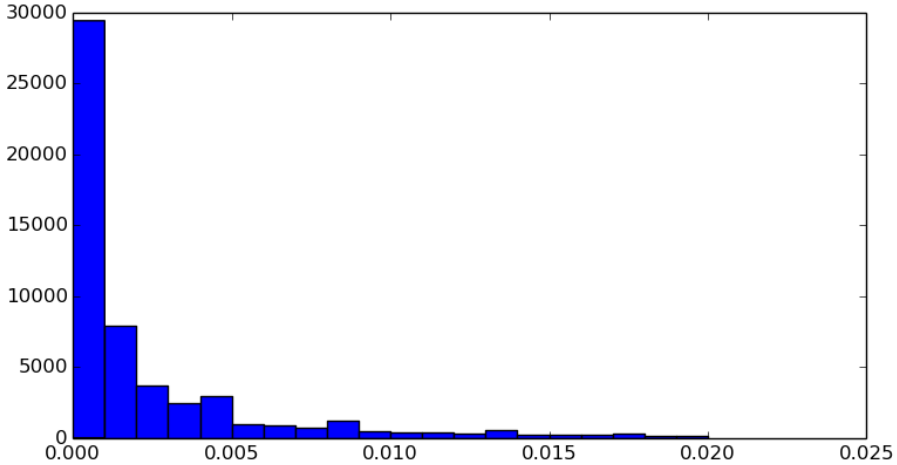


OUT

- **Расстояние до главной страницы в кликах. Полный список приведен в отдельном файле. Гистограмма:**



- **Частота каждого слова приведена в отдельном документе. Гистограмма частоты слов:**



- **Матрица PageRank приведена в отдельном файле. ТОП-20 результатов:**

```
1 http://simple.wikipedia.org/wiki/Main_Page 59.9383240857
2 http://simple.wikipedia.org/wiki/Multimedia 13.6821596634
3 http://simple.wikipedia.org/wiki/United_States 13.1321949991
4 http://simple.wikipedia.org/wiki/France 7.9134416219
5 http://simple.wikipedia.org/wiki/International_Standard_Book_Number 7.40917972814
6 http://simple.wikipedia.org/wiki/Country 5.91078160061
7 http://simple.wikipedia.org/wiki/English_language 5.50790574078
8 http://simple.wikipedia.org/wiki/Canada 5.37596057056
9 http://simple.wikipedia.org/wiki/Information 5.36956408254
10 http://simple.wikipedia.org/wiki/Earth 5.24746189849
11 http://simple.wikipedia.org/wiki/Music 4.96846700488
12 http://simple.wikipedia.org/wiki/India 4.93107299236
13 http://simple.wikipedia.org/wiki/Australia 4.89713755989
14 http://simple.wikipedia.org/wiki/Definition 4.86732515439
15 http://simple.wikipedia.org/wiki/China 4.61581512256
16 http://simple.wikipedia.org/wiki/Sun 4.5325111759
17 http://simple.wikipedia.org/wiki/Islam 4.34192239049
18 http://simple.wikipedia.org/wiki/Christianity 4.32198128908
19 http://simple.wikipedia.org/wiki/Economics 4.2722268045
20 http://simple.wikipedia.org/wiki/Computer 3.97819266155
```

Вывод:

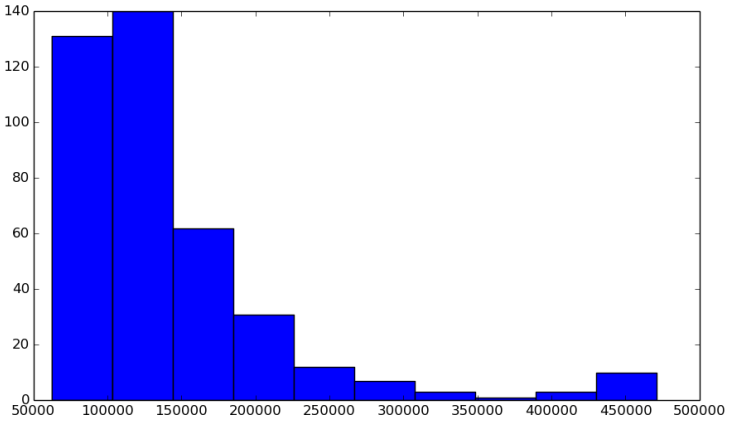
Данный объем скачанной информации с трудом можно назвать полным. Так как 1000 это достаточно небольшая часть всей simple.wikipedia. И полученные результаты очень подвержены суткам, когда они были получены(так как информация на главной страницы меняется). Максимальное расстояния ожидаемое не превышает 2. Так как уже первые 50-70 страниц содержат в сумме более чем 1000 ссылок на другие страницы. Распределения входящих и исходящих степеней вершин похожи между собой. И имеют два экстремума, что объясняется наличием двух групп страниц(маленькие заметки, большие статьи). Последняя содержит большое количество как входящих так и исходящих ссылок. И именно страницы из второй группы имеют больше шансов попасть в топ по PageRank . Гистограмма частоты слов в документе вполне обосновано имеет обратную экспоненциальную зависимость. Присутствуют некоторые шумовые эффекты из-за не столь большой размерности выборки. Весьма очевидно что первой в топ по Page Rank будет начальная страница. Остальные же результаты можно назвать правдоподобными (учитывая ограниченность выборки все-же), по крайне мере некоторые результаты совпадают с оф. приведенной статистикой по всей Вики.

P.S. Оценка итеративного алгоритма PageRank в данном случае O(N^3*log(iter)) или же O(N^2*iter) Что является наилучшим вариантом для некой данной матрицы переходов. Но тем не менее, в данном случае (на практике) количество ребер в ссылочном графе будет линейным или почти линейным. Тогда можно привести алгоритм имеющий время работы O(E*iter) ~ O(N*iter) что уже существенно лучше. И даже может быть посчитано для 10000.

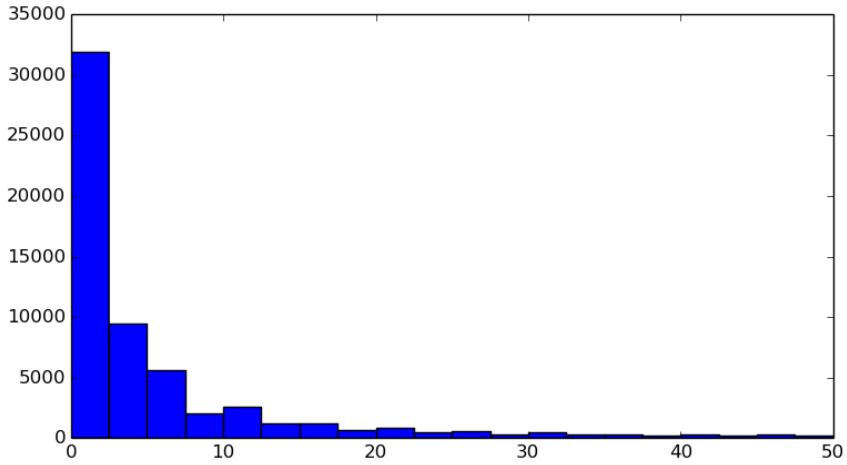
СКАЧИВАЮТСЯ ПЕРВЫЕ 10000 СТРАНИЦ ВИКИПЕДИИ.

(Дополнительная версия, для анализа большого количества скачанных страниц)

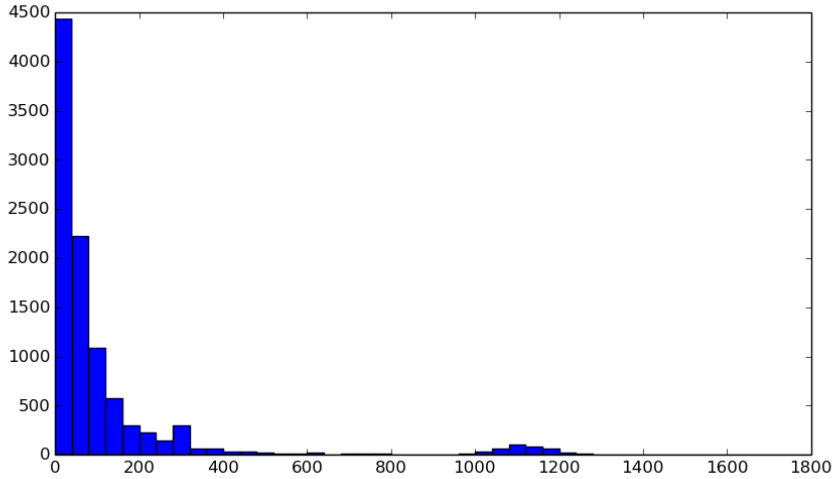
- **Количество рассматриваемых алгоритмов страниц ограничено числов 10000. Данные скачаны 13.04. Размер очереди не ограничивается. На момент обработанных 10000 страниц очередь содержала ~ 50К.**
  Таким образом данные по расстоянию, входящие степени были получены по выборке из 60К страниц.
- **В каждом файле "n.txt" сохраняется по 25 страниц. Средний вес одного файла 200KB. Общий вес ~ 50MB. Время работы алгоритма немного более часа.**
- **Гистограмма распределения размера текстовых документов(байт):**



- **Гистограмма распределения in/out степеней вершин ссылочного графа:**

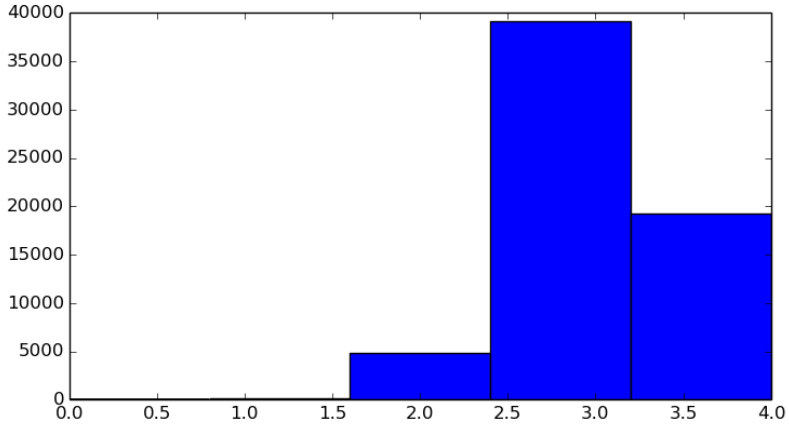


IN

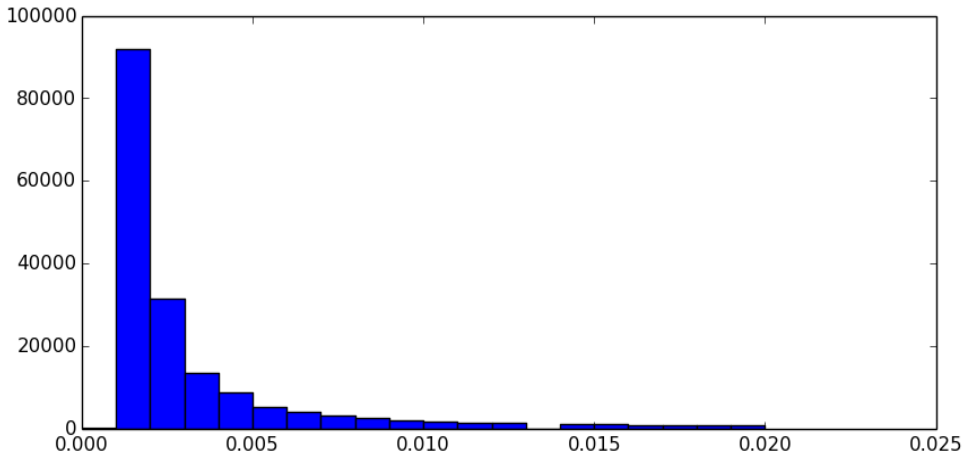


OUT

- **Расстояние до главной страницы в кликах. Полный список приведен в отдельном файле. Гистограмма:**



- **Частота каждого слова приведена в отдельном документе. Гистограмма частоты слов(на коэф):**



Вывод:

Расстояния не превышает 4. Количество страниц в очереди ожидания на загрузку растет достаточно быстро. Достаточно, что бы любая из первых 60 000 была достижима из начальной не более чем за 4 клика. Распределение входящих(In) более пологая и имеет только один очевидный экстремум(при маленьких значениях). Более того график нельзя считать полным. Ведь данные для него получены для 60К, но при рассмотрении 10К страниц, то есть граф не полный. Напротив граф исходящих(Out) степеней построен для 10К страниц(но тем не менее учитываются переходы в первые 60К). То есть этот график можно считать уменьшением реальной полной гистограммы для 60К. На нем так же имеем 2 экстремума. Выделяется группа с большим количеством выходящих ссылок. Гистограмма частоты слов есть обратная экспоненциальная зависимость, что впрочем, неудевительно: слов с низкой частотой не в разы больше часто-встречаемых. Самое встречаемое слово имеет частоту 3.91%. Это “the”. ТОП по Page Ranky в данном случае не приводится, ввиду невозможности его подсчета за разумное время в рамках написанной реализации.(Но выше был предложен способ как можно было бы сделать это возможным, изменив реализацию подсчета Page Ranka)