Домашнее задание 2

В этом задании вам предстоит попрактиковаться в обходе веба с помощью робота и проанализировать граф веб-страниц

1 Задания

Каждое задание оценивается в 2 балла

1. Скачайте подграф Википедии размером порядка 10^4 страниц. Всю информацию о страницах хранить не нужно, достаточно заголовка, исходящих ссылок на другие страницы Википедии и небольшого текстового описания каждой страницы (об этом ниже). При обходе нужно игнорировать ссылк на служебные страницы (File:, Talk:, Category:, Template:, Wikipedia: и пр.), а также Main Page. Рекомендуется использовать фреймворк Scrapy.

   Начать можно с этого примера , а также с кода туториала `http://pybae.github.io/blog/2015/04/27/a-simple-introduction-to-scrapy/`. Следует рассматривать только ссылки из mw-body-content, желательно ограничиваться не более, чем 100 первыми ссылками. Для этого в LinkExtractor можно указать restrict_xpaths или restrict_css - xpath будет выглядеть примерно так:

   ```
   (//div[@id="mw-content-text"]/*/a/@href)[position() < 100])
   ```

   В качестве start_urls выберите произвольный набор из 5 страниц со статьями. Результат выполнения этого задания - файл произвольного формата с информацией о скачанных страницах, с которым вам будет удобно работать дальше.

2. Исходя из сохранённой информации о ссылках между статьями, постройте граф. Рекомендую библиотку networkx

3. Посчитайте PageRank для вершин полученного графа с помощью одной из доступных библиотечных функций. Выведите топ-10 страниц по убыванию PageRank в формате

- *Title <pagerank>*
- *URL*
- *Snippet*

Таким образом, ваш вывод будет похож на настоящую поисковую выдачу.

Что взять в качестве Snippet? Можете решать сами, предлагаю такой вариант:

```
BeautifulSoup(response.xpath('//div[@id="mw-content-text"]/
p[1]').extract_first(), "lxml").text[:255]+"..."
```

Для `https://en.wikipedia.org/wiki/Information_retrieval` получится 'Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the scie...'

4. Поварьируйте alpha, по умолчанию равный 0.85. Попробуйте значения 0.95, 0.5, 0.3. Как изменяется выводимый топ?

5. Примените к полученному графу алгоритм HITS (например, отсюда `https://networkx.github.io/documentation/networkx-1.9.1/reference/generated/networkx.algorithms.link_analysis.hits_alg.hits.html`). Сравните топ-10 результатов по значениям авторитетности, хабовости и их среднего с топом по PageRank. Сделайте выводы о наблюдаемых различиях.

   Примечание: HITS действительно используется для анализа взаимосвязей между потенциально релевантными запросу документами, а здесь никаких и запросов-то нет. Смысл задания в том, чтобы сравнить способ оценки "важности"страницы с помощью одного параметра (как в PageRank) и с помощью разных параметров HITS.

2 Требования

В качестве решения принимается код, сжатый архив страниц Википедии и отчёт (pdf / ipynb). По этой задаче soft- и hard- дедлайны совпадают.