Лаб. Работа № 2

Получение и подготовка исходных данных для анализа из открытых источников, не требующих авторизации и не имеющих API

**Цель работы** – приобрести навыки работы с исходными данными, получаемыми из открытых источников сети Интернет

**Задание:** получить на вход программы URL страницы интернет-ресурса, а так же параметр *depth* - уровень глубины исследования ссылок, присутствующих на заданной странице.

C помощью пакетов [scrapy](https://scrapy.org/) (уже присутствует в Anaconda), либо **[BeatifulSoup](http://www.crummy.com/software/BeautifulSoup/bs4/doc/)**, **[lxml](http://lxml.de/index.html)** плюс пакета для оформления запросов - [Requests](http://docs.python-requests.org/en/master/) осуществить просмотр/парсинг заданной веб-страницы и собрать статистику следующего вида:

1. Количество ссылок на странице анализируемого уровня
2. Количество слов/символов на странице анализируемого уровня
3. Построить гистограмму частоты встречаемости слов на странице анализируемого уровня.
4. Построить гистограмму встречаемости длин слов (т.е. сколько слов состоит из 1,2,3,4… и т.п. k- символов).
5. Построить таблицу ссылок и для каждой ссылки из таблицы повторить шаги 1-4.
6. Объединить статистику разных страниц в соответствующие таблицы (ш. 1-4), отображающие единую статистику по всем проанализированным страницам.

**Доп. варианты получаемой статистики:**

- количество изображений на странице

- общий и средний размер изображений на странице

- количество заголовков и подзаголовков (- в зависимости от тегов заголовков разных уровней)

- гистограмма встречаемости длин слов в подзаголовках

*- добавить что-то ещё на собственный выбор*

**Доп. задание на бонус:**

- оформить облако тегов по странице и ссылкам

**Информация для подготовки:**

- <https://habr.com/ru/post/280238/>

- <https://proglib.io/p/parsing-course/>

- <http://doc.scrapy.org/en/latest/intro/overview.html> (English)