Offliner

December 20, 2015

# Описание

Программа Offliner предназначена для чтения оффлайн веб-страниц в текстовом формате. Для этого программе сначала подается требуемый веб-адрес (url). Эта страница выкачивается, вместе с попутными картинками и сохраняется на диске в директории с названием корневого домена со знаком подчеркивания. Например, головная страница новостной системы lenta.ru будет сохранена в директории \_lenta.ru

# Архитектура

Программа Offliner разработана на языке Python и имеет следующую архитектуру:

**Родительский класс HTMLParser из стандартной библиотеки html.parser**

HTML content

**Основной класс Offliner**

**Описание**

* Обработка входных параметров
* Извлечение исходного HTML документа или изображения
* Сохранение результатов

**Методы**

* **process\_data()** – головной метод, точка входа - содержит всю логику программы
* **\_retrieve\_data()** – служебный метод для выкачивания веб-документа или изображения.
* **\_get\_charset ()** – служебный метод для преобразования charset в формат, понимаемый интерпретатором python

Text content, image URLs

**Наследуемый класс Http**

**Описание**

* Обработка значащих html тегов
* Удаление незначащих тегов вроде <script>
* Возвращает текст и список изображений

**Методы**

* **handle\_starttag()** – обработка стартовых тегов <tag>
* **handle\_endtag()** – обработка финишных тегов </tag>
* **handle\_data()** – обработка текста между тегами <tag>текст</tag>
* **get\_content() –** возвращает веб-страницу в текстовом формате и список изображений для выкачивания в формате полных веб-адресов, разделенных пробелами

# Условные обозначения в выходном текстовом формате

В результате работы программы получается текстовый файл, в котором:

* в квадратных скобках [] обозначаются ссылки на изображения и другие страницы
* в треугольных скобках <> обозначаются оригинальные имена изображений
* звездочками обозначаются пункты именованных и неименованных HTML списков <ul>

# Дальнейшее развитие программы

Ниже перечислены текущие ограничения для устранения и пути дальнейшего развития программы

* Возможность задавать директорию с выкачанными страницами, специфицировать имя выходного файла
* Работа с некорректными/нестандартными страницами, где например
  + Нет тегов <body>, <title>
  + Не задан в ответе http-сервера charset с кодировкой документа. В таком случае программа считает кодировку utf-8 по умолчанию. Корректно анализировать тег <meta>:

<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

* Корректная работа с тагами <base>, <iframe>
* Улучшение форматирования текстового файла, избавление от незначащих/служебных элементов
* Логирование и отладочный режим, включаемый по опции –verbose для отладки
* Работа со всеми тегами, документированными спецификацией html
* Работа со страницами, требующими идентификации пользователя