Разработка парсера-менеджера

Выполнил: Светлов Даниил

Постановка задачи

В предложенном задании было необходимо выполнить 2 задачи:

- разработка и тестирование парсера web-страниц;
- разработка и тестирование парсера документов.

Оценка времени

	Позиция	Задачи	Затраточасы
Светлов Даниил	Менеджер	Распределение обязанностей, подготовка отчетной документации, создание продуманной архитектуры	5-8
Светлов Даниил	Разработчик	Согласование архитектуры, написание кода программы, исправление багов, выявленных в процессе тестирования	8-10
Светлов Даниил	Тестировщик	Определения тест-кейсов, написание тестов, выявление багов программы	4-5

Архитектура

В связи с отсутствием жестко регламентированного технического задания было принято решение произвести синтаксический анализ для:

- страниц сайтов СПбГУ и МГУ;
- файлов расширения “.docx”, “.pdf”, содержащих текст на различных языках и таблицы.

Результаты сохраняются локально в формате “.txt”.

Стек технологий

Реализация выполнена на языке python. Данный выбор был сделан на основе следующих суждений:

- основной язык разработки команды;
- имеется достаточное кол-во библиотек, позволяющих сократить время разработки.

Основные зависимости

- bs4(библиотека, упрощающая сбор информации с веб-страниц);
- pdfminer(инструмент для извлечения текста из PDF-документов);
- docx2txt(утилита для извлечения текста из файлов docx).

Тестирование

Были прописаны следующие тест-кейсы:

- файлы “.docx”, “.pdf” на английском и русских языках;
- пустые файлы;
- файлы со спец. символами;
- многостраничные файлы;
- веб страницы.

В процессе тестирования багов не обнаружено.

Итоги

Продукт на выходе имеет следующий функционал:

- возможность парсить страницы сайтов СПбГУ и МГУ;
- парсить документы формата “.docx”, “.pdf” содержащие текст и таблицы.

Данные пункты удовлетворяют всем критериям поставленной задачи.

Разработка поискового робота для сбора и обработки данных с ресурсов Web 1.0

Постановка задачи

В предложенном задании было необходимо выполнить 3 задачи:

- разработка модели поискового робота для сбора и обработки данных в сети Web 1.0;
- автоматизированный сбор данных с помощью робота на сайтах СПбГУ и МГУ;
- сбор статистики обработанных страниц.

Оценка времени

	Позиция	Задачи	Затраточасы
Светлов Даниил	Менеджер	Распределение обязанностей, подготовка отчетной документации, создание продуманной архитектуры	5-8
Светлов Даниил	Разработчик	Согласование архитектуры, написание кода программы, мониторинг работы программы, сбор результатов(*)	10-12 * - 3 дня
Светлов Даниил	Тестировщик	-	0

Архитектура

В качестве основного языка разработки был выбран python.

Также были задействованы следующие зависимости:

- scrapy(сканирование веб-сайтов и извлечение структурированных данных с их страниц)

Тестирование

Были прописаны следующие тест-кейсы:

- Простая страница с текстом
- Страница с текстом на русском
- Страница с картинками, списками, кнопками
- Простая ссылка
- Рекурсивная ссылка
- Последовательные ссылки для проверки правильности глубины поиска
- Страница со сломанной ссылкой
- Автоматическая переадресация переадресация через время

В процессе тестирования багов не обнаружено.

Статистика

На основе собранных и обработанных данных была собрана следующая статистика:

Для сайта СПбГУ:

- общее количество внутренних ссылок - 11575831 ;
- общее количество страниц - 108857;
- количество внутренних страниц - 103160;
- количество неработающих страниц - 31(404), 77 (403);
- количество внутренних поддоменов - 208;
- количество уникальных внешних ресурсов - 3305.

Статистика

Для сайта МГУ:

- общее количество внутренних ссылок - 4728906 ;
- общее количество страниц - 95073;
- количество внутренних страниц - 79483;
- количество неработающих страниц - 19(404), 277 (403);
- количество внутренних поддоменов -392;
- количество уникальных внешних ресурсов - 8385.

Итоги

Продукт на выходе имеет следующий функционал:

- имеет модель поискового робота для сбора и обработки данных в сети Web 1.0;
- автоматизированный сбор данных с помощью робота на сайтах СПбГУ и МГУ;
- сбор статистики обработанных страниц.

Данные пункты удовлетворяют всем критериям поставленной задачи.