Парсинг HTML-документов

Подготовлено для: студентов группы 149

Подготовлено: Кирпичёв А.Н.

16 марта 2017 г.

Номер работы: 3

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цель

Научится получать информацию с различных web-сайтов посредствам специальных классов

Задачи

- Изучить структуру страниц для распознания
- Выбрать оптимальное средство парсинга для решения задачи
- Изучить класс
- Применить класс для поставленной задачи

Отчет

- Разработанное приложение

Задание

- Создать необходимые таблицы для хранения данных
- Скрипт должен работать для любой подобной страницы
- Выбор страниц поместить в таблицы базы данных
- Во время защиты работы уметь модифицировать работу парсера.

Инструменты

- https://github.com/hxseven/htmlSQL - htmlSQL
- https://code.google.com/p/phpquery/ -phpquery
- http://querypath.org/ - QueryPath
- https://github.com/amal/CDom CDom
- http://sourceforge.net/projects/simplehtmldom/ - PHP Simple HTML DOM Parser

ПЕРЕЧЕНЬ ТЕМ

№	Пример страницы	Извлекаем
1	http://www.kinopoisk.ru/film/71065/	Данные о произвольном фильме, включая информацию об актерах (с их фото)
2	http://market.yandex.ru/product/7746371/	Данные о блоке питания, о обзорах и отзывах
3	http://www.asus.com/ru/Motherboards/GRYPHON_Z97_ARMOR_EDITION	Получать полную информацию о продукте
4	http://www.softclub.ru/games/pc/22587-vedmak-3-dikaya-okhota	Получение полной информации об игре
5	http://www.rlsnet.ru/tn_index_id_1578.htm	Полные данные о лекарсивенном средстве
6	http://www.imdb.com/title/tt3235888/?ref_=hm_cht_t4	Данные о произвольном фильме, включая информацию об актерах (с их фото)
7	http://www.mvideo.ru/products/televizor-philips-ultrahd-55pus7809-60-10006417	Получать полную информацию о продукте
8	http://info.tmgame.ru/library/world/bestiary/5lvl/	Распознать информацию о NPC
9	http://royalquest-db.ru/predmetyi/97-abordazhnaja_sablja.html	Распознавать информацию о предметах
10	http://www.dns-shop.ru/catalog/i1007387/156-noutbuk-dexp-atlas-h104-seryj	Загружать информацию о товаре, его стоимость и отзывы о товаре
11	https://tv.yandex.ru/19/channels/146	Загружать данные о телепрограмме
12	http://www.cbr.ru/hd_base/Default.aspx?Prtid=metall_base_new	вытаскивать данные по драг металлам за любой период
13	http://www.cbr.ru/currency_base/daily.aspx?date_req=01.03.2015	курс валют на любую дату
14	https://toster.ru/q/206698	Организовать парсинг ответов на вопрос
15	https://pogoda.mail.ru/prognoz/syktyvkar/1-april/#2014	Распознавать данные о погоде
16	https://hi-tech.mail.ru/lenovo_sisley_s90_14230324-catalog/	Получать полную информацию о продукте

№	Пример страницы	Извлекаем
17	https://afisha.mail.ru/cinema/movies/721723_patrul_vremeni/	Данные о произвольном фильме, включая информацию об актерах
18	https://tv.mail.ru/channel/1395/117/	Загружать данные о телепрограмме
19	https://slovari.yandex.ru/%D1%81%D0%BB%D0%BE%D0%BD/%D0%BF%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81%D0%B0%D0%BD%D0%B8%D0%B5/	Загружать информацию о слове
20	http://quote.rbc.ru/cash/bank/63493.html	Извлекать информацию о банке и о курсах валют данного банка
21	http://sochi.lenta.ru/countries/russia/index.html	Уметь извлекать информацию о результатах и медалях спортсменов стран
22	http://sochi.lenta.ru/sports/snowboarding/index.html	Уметь обрабатывать результаты по любому виду спорта
23	https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B8%D1%81%D0%BE%D0%BA_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2%D1%8B%D1%85_%D1%80%D0%B5%D0%B4%D0%B0%D0%BA%D1%82%D0%BE%D1%80%D0%BE%D0%B2	Уметь вытаскивать данные из таблиц сравнения на любых страницах wiki
24	http://bash.im/index/1020	Уметь распознавать записи на странице
25	http://www.3dnews.ru/220013	Уметь распознавать информацию о любом приложении

РАСПРЕДЕЛЕНИЕ ТЕМ

Группа	НОмер темы
Бастанжиева О.	10
Баум В.	17
Габов И.	16
Грабовский А.	14
Калкина Е.	9
Комаров П.	8
Михайлова Т.	11
Хвищук Е.	4
Цветкова Т.	18
Шилова Л.	3