Санкт-Петербургский Государственный Университет
Кафедра системного программирования

Загрузка и хранение информации для приложения “Дом вещей”

Кудряшова Анна Александровна
Научный руководитель: ст. преп. Баклановский М.В.

2017

Введение

Загрузчик

NFC метка

Телефон на базе Android

Android приложение

Пользователь

Обзор предметной области

Краулер — программа предназначенная для перебора страниц Интернета. Он анализирует содержимое страницы, сохраняет его в некотором специальном виде.

Веб-скрепинг — процесс извлечения информации из интернета. С целью ее хранения и дальнейшего использования.

Краулер является составной частью веб-скрепинга.

Цель

Реализовать получение информации о бытовой и цифровой технике из интернета для предоставления этой информации пользователю через приложение.

Задачи

- сделать обзор методов парсинга и краулинга веб-страниц
- реализовать поиск по ключевым словам средствами поисковых систем
- реализовать парсинг страниц с целью извлечения необходимой информации
- реализовать хранение информации в базе данных
- реализовать скачивание документов через скрипт
- реализовать скачивание изображения

Обзор существующих инструментов

Инструменты для краулинга

Функциональность	BeautifulSoup	Scrapy
Отправка запросов	+ (необходима доп. библиотека Request)	+
Проход по ссылкам в глубину	-	+
Защита от распознания бота	+ (необходимо настраивать самостоятельно)	+ (настраивается автоматически)
Загрузка изображений и pdf файлов	+ (необходима доп. библиотека)	+

Обзор существующих инструментов

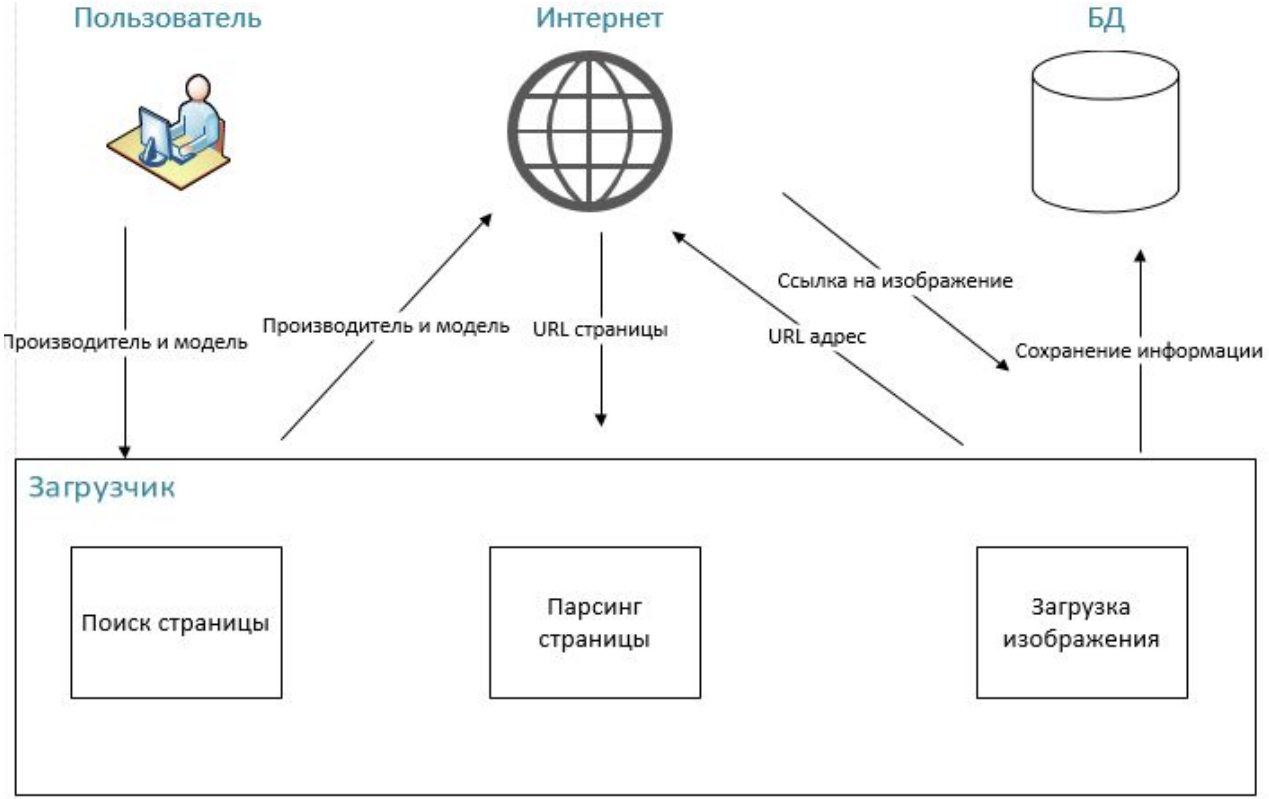
Инструменты для хранения информации

Функциональность	MongoDB	SQLite+Sqlalchemy
Поддержка кириллицы	+	+
Использование на android	-	+
База представлена одним файлом	-	+
Быстрый отклик	+	+

Инструменты

- python
- framework scrapy
- SQLite + Sqlalchemy
- web-framework Flask

Архитектура решения



Реализация

- **Поиск страницы**

Модель и производитель подставляются в URL адрес в качестве атрибутов

- **Парсинг страницы**

Поиск необходимой информации производится по содержимому текста средствами синтаксиса xpath

- **Скачивание изображения**

Скачивание изображения осуществляется с Яндекс.Маркет

Реализация

- **Скачивание инструкции**

Для загрузки инструкции было рассмотрено несколько возможных источников:

1. Сайт производителя
2. Сайты интернет-магазинов

Реализация

- **Хранение информации**

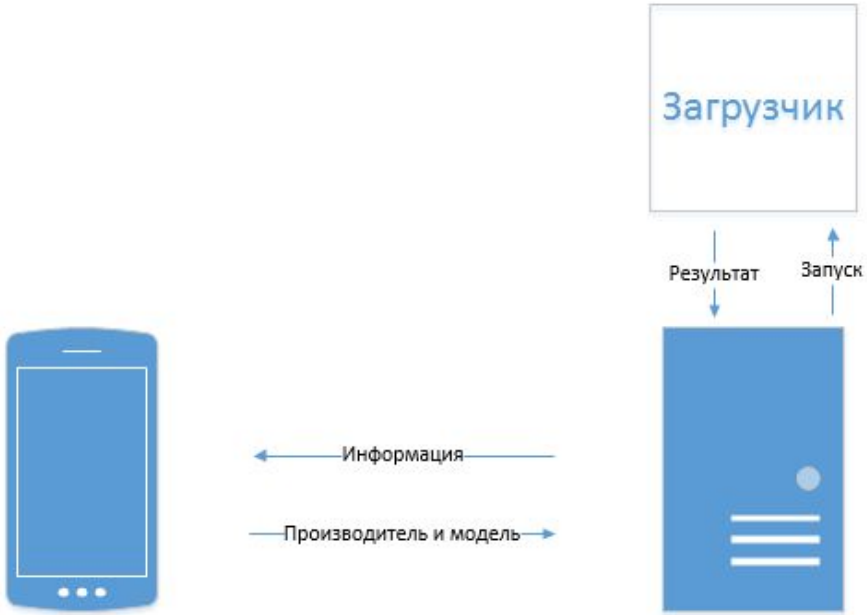
Для хранения информации используется единая для всех типов устройств концепция, которая включает в себя следующие параметры:

- модель
- производитель
- тип устройства
- параметры(ДхШхВ)
- вес
- ссылка на сайт производителя
- инструкция
- магазин

Реализация

- **Веб-сервер**

Архитектура взаимодействия локального сервера и клиента



Результат

- изучена предметная область парсинга сайтов с использованием краулеров
- сделан обзор существующих инструментов для парсинга сайтов на языке Python
- реализован скрипт по поиску, загрузке информации и изображений с html страниц
- изучена работа базы данных SQLite
- реализован скрипт по сохранению найденной информации в базу данных
- реализована работа загрузчика на локальном сервере
- реализовано взаимодействие локального сервера и android приложения