|  |  |
| --- | --- |
|  | |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  Федеральное государственное автономное образовательное учреждение высшего образования  **«Дальневосточный федеральный университет»**  (ДВФУ) | |
|  | |
| **ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ (ШКОЛА)** | |
| **Департамент математического и компьютерного моделирования** | |
| **КУРСОВОЙ ПРОЕКТ** | |
| по дисциплине «Программирование баз данных» | |
| на тему «Создание реляционной базы данных на основе парсинга веб-сайтов (магазины линолеумов)» | |
| по образовательной программе подготовки бакалавров | |
| по направлению 09.03.03 «Прикладная информатика» | |
| профиль Прикладная информатика в компьютерном дизайне | |
| Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Выполнил студент группы  № Б9122-09.03.03пикд  \_\_\_\_\_\_\_\_\_\_\_\_\_ Сакмаркин В. А.  (подпись)  «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024г. |
| Регистрационный номер \_\_\_\_\_\_\_  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2024г. | Руководитель: старший преподаватель  \_\_\_\_\_\_\_\_\_\_\_\_\_ Селезнев Т. Э.  (подпись)  «\_\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2024г. |
| г. Владивосток  2024 | |

# **Оглавление**

[Оглавление 2](#_Toc169540411)

[Введение 3](#_Toc169540412)

[Изучение и выбор методов реализации 4](#_Toc169540413)

[1.1 Обзор существующих методов решения 4](#_Toc169540414)

[1.2 Обоснование выбранных методов и средств реализации 5](#_Toc169540415)

[1.3 Выбор подходящих веб-сайтов и их анализ 5](#_Toc169540416)

[1.4 Возможные проблемы и способы их решения 8](#_Toc169540417)

[Реализация проекта на основе выбранных методов 9](#_Toc169540418)

[2.1 Разработка алгоритмов парсинга 9](#_Toc169540419)

[2.2 Создание базы данных и внесение полученных данных 9](#_Toc169540420)

[2.3 Разработка веб-сайта 9](#_Toc169540421)

[Заключение 10](#_Toc169540422)

[Список литературы 11](#_Toc169540423)

# Введение

В современном информационном обществе инструменты автоматизации все чаще применяются для получения и обработки данных из сети Интернет. Автоматизация этих процессов позволяет рационализировать и оптимизировать работу в области поиска и анализа информации. Для получения и обработки данных из сети Интернет используется парсинг. Парсинг веб-сайтов представляет собой процесс извлечения данных, структуры и характеристик веб-страниц с использованием специализированных инструментов.

Целью курсовой работы является разработка программного решения для парсинга веб-сайтов магазинов линолеума, а также сохранение полученных данных в базу данных. Для удобного просмотра полученных данных будет разработан веб-сайт.

В общем случае, парсинг веб-сайтов применяется во многих сферах: маркетинг и исследование рынка, финансы и инвестиции, научные исследования, реклама и маркетинг. В маркетинге и исследовании рынка парсинг позволяет получить данные о конкурентах, ценах, товарах и услугах, динамике спроса и предложения на рынке, а также облегчает процесс составления прайс-листов магазинов, в случаях парсинга сайтов компаний-поставщиков.

В работе будет приведен обзор существующих методов парсинга веб-сайтов, описан выбор сайтов и анализ их структуры и содержания. Затем будет разработан алгоритм парсинга для каждого выбранного сайта, создана реляционная база данных, в которую будут внесены полученные с веб-сайтов данные. Далее будет разработан веб-сайт, который позволит удобно просматривать информацию из базы данных.

# Изучение и выбор методов реализации

## Обзор существующих методов решения

Существует ряд методов в области парсинга веб-сайтов, которые позволяют извлекать данные с веб-страниц. Ниже приведены те, которые чаще всего используются специалистами парсинга.

1. API сайтов: API – программный интерфейс, который позволяет одной программе взаимодействовать с другой. Некоторые веб-сайты предоставляют официальные API для получения данных. Как правило, такой способ парсинга наиболее легкий, однако официальные API встречаются редко.
2. Получение данных из XHR-запросов: Чаще всего, современные веб-сайты используют JavaScript для загрузки данных с сервера путем отправления GET или POST запросов после получения структуры страницы, что делает загрузку содержимого постепенной и плавной. В инструментах разработчика любого браузера можно посмотреть эти запросы, что дает возможность повторить их в программе для парсинга. Такой способ является удобным и надежным в долгосрочной перспективе, поскольку в случае изменения ответов сервера их структура остается неизменной.
3. Поиск JSON в HTML-коде страницы: Чтобы страница корректно индексировалась поисковыми системами, необходимо, чтобы в HTML-коде страницы содержалась вся нужная информация. При генерации страницы на стороне сервера в HTML-код нередко добавляется JSON (текстовый формат обмена данными, основанный на JavaScript), содержащий данные этой страницы. Этот способ используется в том случае, когда отсутствует возможность получения данных предыдущим методом, однако извлечение JSON может быть трудным.
4. Парсинг HTML-кода: Применяется тогда, когда применение всех вышеупомянутых способов невозможно. Данные извлекаются непосредственно из элементов страницы. Для этого часто используются специальные библиотеки, например, BeautifulSoup4 или Scrapy. Способ подразумевает анализ структуры веб-страницы и извлечение данных из элементов с определенными тэгами или атрибутами. Такой подход является наименее надежным в долгосрочной перспективе, поскольку структура страницы со временем может быть изменена.

## 1.2 Обоснование выбранных методов и средств реализации

Поскольку целью работы не является разработка программного решения, которое будет надежным в долгосрочной перспективе, было решено использовать парсинг HTML-кода в качестве метода извлечения данных. Для этого была выбрана библиотека BeautifulSoup4 для языка программирования Python, так как она проста в использовании и имеет понятную официальную документацию.

Для разработки веб-сайта выбран фреймворк Django, реализуемый на языке Python. Стоит отметить, что данный фреймворк имеет специфические методы работы с базами данных. Об этих методах будет подробно изложено в пункте «Создание базы данных и внесение полученных данных» главы «Реализация проекта на основе выбранных методов».

## 1.3 Выбор подходящих веб-сайтов и их анализ

Для выполнения работы были выбраны следующие сайты магазинов, продающих линолеум:

**Топтыгин**

Топтыгин (<https://polov.net>) – магазин напольных покрытий и сопутствующих товаров. Магазин предлагает материалы самых популярных брендов, что делает его выбор обоснованным.

Анализ структуры страниц:

* Наименование товара: Расположено в тэге <h1>. Атрибут class учитывать необязательно, поскольку по правилам верстки HTML-страниц тэг <h1> на странице может быть только один.
* Цена товара: Расположена в тэге <span>, вложенном в тэг <div> с классом “product-item-detail-unit-price”.
* Характеристики товара: Расположены в тэгах <li> с классом “product-item-detail-properties-item”, где имя характеристики находится в тэге <span> с классом “product-item-detail-properties-name”, а значение характеристики – в тэге <span> с классом “product-item-detail-properties-value”. К характеристикам товара относятся: ширина, толщина, толщина защитного слоя, класс пожарной безопасности и бренд.
* Ссылка на изображение: Расположена в атрибуте src тэга <img>, вложенном в тэг <div> с классом “product-item-detail-slider-image”.

**Залог**

Залог (<https://zalog-vostok.ru>) – магазин строительных товаров во Владивостоке. Магазин предлагает широкий ассортимент товаров, в том числе и линолеумов, что делает его хорошим вариантом выбора.

Анализ структуры страниц:

* Наименование товара: Расположено в тэге <h1>.
* Цена товара: Расположена в тэге <span> с классом “price\_value”.
* Характеристики товара: Расположены в тэгах <div> с классом “properties\_\_item”, где имя характеристики находится в тэге <div> с классом “properties\_\_title”, а значение характеристики – в тэге <div> с классом “properties\_\_value”. К характеристикам товара относятся: ширина, толщина, толщина защитного слоя, класс пожарной безопасности и бренд.
* Ссылка на изображение: Расположена в атрибуте src тэга <img>, с классом “detail-gallery-big\_\_picture”.

**ПолДома**

ПолДома (<https://vladivostok.pol-doma.com>) – магазин напольных покрытий в крупных городах России. Магазин предлагает широкий ассортимент товаров различных брендов и ценовых категорий (в частности, более 2000 видов линолеума), что делает его достаточно ценным ресурсом для извлечения данных.

Анализ структуры страниц:

* Наименование товара: Расположено в тэге <h1>.
* Цена товара: Расположена в тэге <span> с классом “price\_value”. В отличие от предыдущих сайтов, цена указана за один квадратный метр линолеума, а не за один погонный метр.
* Характеристики товара: Расположены в тэгах строк таблицы <tr>, где имя характеристики находится в тэге <td> с классом “char\_name”, а значение характеристики – в тэге <td> с классом “char\_value”. К характеристикам товара относятся: ширина, толщина, толщина защитного слоя и класс пожарной безопасности.
* Бренд: Содержится непосредственно в наименовании товара и заключен в кавычки.
* Ссылка на изображение: Расположена в атрибуте href тэга <a>, с классом “popup\_link”.

Выбор этих сайтов обусловлен их популярностью, доступностью, наличием необходимых характеристик товаров, а также разнообразием товаров.

## 1.4 Возможные проблемы и способы их решения

Извлечение данных о товарах из каталога магазина подразумевает собой обращение к серверу для получения веб-страницы каждого товара. Однако слишком частые запросы из одного источника сервер может посчитать за подозрительный трафик и заблокировать этому источнику доступ по его IP-адресу.

Решений у этой проблемы несколько:

1. Как правило, такие блокировки ограничены по времени, а значит можно дождаться ее завершения.
2. Если у используемого Интернет-провайдера не подключена услуга «Статический IP-адрес», можно дождаться смены адреса.
3. Чтобы не ждать, можно использовать несколько прокси-серверов и переключаться между ними в случае блокировок.

Наилучшим вариантом является предотвращение появления данной проблемы. Например, можно изначально использовать прокси-сервера и переключаться между ними с определенной периодичностью или же можно делать паузы между запросами на сервер длительностью в несколько секунд.

# Реализация проекта на основе выбранных методов

## 2.1 Разработка алгоритмов парсинга

## 2.2 Создание базы данных и внесение полученных данных

## 2.3 Разработка веб-сайта

# Заключение

В ходе работы над курсовым проектом на игровом движке Unity была разработана 3D игра в жанре шутер от первого лица. Была реализована одна из основных механик игры – механика стрельбы. А также был реализован игровой интерфейс, в том числе главное меню и меню паузы.

В ходе работы возникли различные проблемы. Главной из них стал сбой в работе операционной системы, из-за чего понадобилась ее переустановка. Также возникали конфликты между программными модулями из-за неправильной реализации их взаимодействия.

Работа над проектом дала опыт работы с игровым движком Unity, а также опыт работы в команде – один из важнейших навыков в профессиональной деятельности.

# Список литературы

1. Newzoo Global Games Market Report 2022 | Newzoo [Электронный ресурс] / URL: <https://newzoo.com/resources/trend-reports/newzoo-global-games-market-report-2022-free-version>
2. Руководство Unity | Unity [Электронный ресурс] / URL: <https://docs.unity3d.com/ru/530/Manual/UnityManual.html>
3. Статья «Что такое Raycast в Unity и как он работает?» | Stack Overflow [Электронный ресурс] / URL: <https://ru.stackoverflow.com/questions/845129/%D0%A7%D1%82%D0%BE-%D1%82%D0%B0%D0%BA%D0%BE%D0%B5-raycast-%D0%B2-unity-%D0%B8-%D0%BA%D0%B0%D0%BA-%D0%BE%D0%BD-%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0%D0%B5%D1%82>
4. Статья «Как работает формула поворота вектора на угол?» | Stack Overflow [Электронный ресурс] / URL: <https://ru.stackoverflow.com/questions/1409394/%D0%9A%D0%B0%D0%BA-%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0%D0%B5%D1%82-%D1%84%D0%BE%D1%80%D0%BC%D1%83%D0%BB%D0%B0-%D0%BF%D0%BE%D0%B2%D0%BE%D1%80%D0%BE%D1%82%D0%B0-%D0%B2%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%B0-%D0%BD%D0%B0-%D1%83%D0%B3%D0%BE%D0%BB>