Федеральное государственное автономное

образовательное учреждение

высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |
| --- |
| Институт космических и информационных технологий |
| *институт* |
| Вычислительная техника |
| *кафедра* |

**ОТЧЕТ О ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ**

|  |
| --- |
| **КрасИнсайт** |
| *место прохождения практики*  **Автоматизация браузера для сбора данных с использованием Puppeter** |
| *тема* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Руководитель от университета | | |  |  |  | М.В. Савицкая |
|  | |  |  | *подпись, дата* |  | *инициалы, фамилия* |
| Руководитель от предприятия | | |  |  |  | М.С. Медведев |
|  |  | |  | *подпись, дата* |  | *инициалы, фамилия* |
| Студент | КИ19-09Б, №031940750 | |  |  |  | Н.М. Кудрявцев |
|  | *номер группы, зачетной книжки* | |  | *подпись, дата* |  | *инициалы, фамилия* |

Красноярск 2022 г.

**ОГЛАВЛЕНИЕ**

[**ВВЕДЕНИЕ** 3](#_Toc108950664)

[1. Характеристика предприятия: 4](#_Toc108950665)

[2. Этапы прохождения производственной практики: 5](#_Toc108950666)

[3. Задание для прохождения производственной практики: 6](#_Toc108950667)

[4. Подготовка к выполнению задания: 6](#_Toc108950668)

[5. Описание выполнения задания по производственной практике: 6](#_Toc108950669)

[**ЗАКЛЮЧЕНИЕ** 12](#_Toc108950670)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ** 13](#_Toc108950671)

# **ВВЕДЕНИЕ**

Производственная практика «Технологическая (проектно-технологическая) практика» является одним из компонентов образовательной программы и предназначена для проведения практической подготовки. Практическая подготовка – это форма организации образовательной деятельности при освоении образовательной программы в условиях выполнения обучающимися определенных видов работ, связанных с будущей профессиональной деятельностью и направленных на формирование, закрепление, развитие практических навыков и компетенций.

Данная производственная практика проходила в компании «КрасИнсайт» в соответствии с графиком учебного процесса с 22.06.2022 по 19.07.2022.

## 1. Характеристика предприятия:

Сибирский федеральный университет — многопрофильный научно-образовательный центр, крупнейший университет Восточной Сибири [создан](http://about.sfu-kras.ru/history) в 2006 году объединением четырех крупнейших вузов Красноярского края.

СФУ относится к классу «быстроразвивающихся молодых университетов» и является единственным представителем российской высшей школы среди лучших мировых университетов-миллениалов по версии Times Higher Education в 2020 году, входит в топ-350 рейтинга лучших университетов мира в возрасте до 50 лет в 2021 году, а также занимает 122-е место в мире и 2-е среди вузов России в глобальном рейтинге устойчивого развития и качественной университетской среды UI GreenMetric в 2021 году. [Подробнее](http://about.sfu-kras.ru/rating)

СФУ является крупнейшим хозяйствующим субъектом и работодателем региона и входит в перечень cистемообразующих предприятий в сфере образования. Ежегодный набор на первый курс составляет более 5,3 тыс. студентов, общее количество обучающихся на конец 2020 года составило более 26 тыс. человек. С 2007 года дипломы Сибирского федерального университета получили более 100 тыс. человек. Организационная [структура](http://structure.sfu-kras.ru/departments) университета отражает его масштаб и разнообразие: 22 профильных института, 3 филиала, 6 научно-образовательных центров, пояс R&D-центров, 108 лабораторий, 32 базовые кафедры.

Сегодня университет интегрирован практически во все процессы технологического, экономического, социального и культурного развития Красноярского края и Сибирского макрорегиона и оказывает значительное влияние на их динамику.

На новом этапе развития университета ключевой целью станет обеспечение опережающего социально-экономического развития Ангаро-Енисейского макрорегиона.

В качестве места прохождения производственной практики была выбрана кафедра ВТ ИКИТ СФУ.

Кафедра Вычислительной техники была организована в 1969 году в Красноярском политехническом институте (КГТУ на момент создания СФУ). За более чем полувековую историю кафедры подготовка специалистов осуществлялась на всех поколениях вычислительной техники от отечественных персональных цифровых вычислительных машин для инженерных расчетов (МИР) до самого современного оборудования.

С 1983 года кафедра стала выпускающей. С первого выпуска, состоявшегося в 1988 году, было подготовлено и выпущено более 2000 выпускников (специалистов, бакалавров, магистров). Основной коллектив кафедры составляют ее выпускники.

Кафедра сотрудничает с Министерством обороны РФ, Министерством здравоохранения РФ, Министерством науки и высшего образования РФ, Министерством экономического развития РФ, ведущими предприятиями и организациями России: ОАО «ИСС», РУСАЛ, РЖД, банковская сфера и др., а также со многими университетами России и других стран.

Выпускники кафедры занимают руководящие и ведущие должности во всех отраслях страны и за рубежом. География мест работы выпускников – Россия, Европа, Азия, Америка.

В структуру кафедры входят научно-учебные лаборатории:  [Микропроцессорных систем](http://ikit.sfu-kras.ru/s/55), [Телекоммуникационных систем](http://ikit.sfu-kras.ru/s/29),  [Технологии программирования](http://ikit.sfu-kras.ru/s/44),  [Систем автоматизированного проектирования](http://ikit.sfu-kras.ru/s/66)

**Компания «КрасИнсайт»** занимается созданием сайтов в Красноярске более 15 лет. Компания готова предложить свои услуги по разработке, поддержке и продвижению веб-сайтов. Их портфолио включает проекты известных компаний Красноярска и России. Также они могут помочь решить все вопросы, связанные с дальнейшей технической поддержкой и обслуживанием. На каждый проект предоставляется гарантия 5 лет.

Стоит отметить, что любой этап работ (от разработки фирменного стиля, дизайна, программирования, информационного наполнения до поискового продвижения) находится под контролем команды штатных специалистов, это позволяет получить качественный продукт, который поможет вам достичь поставленных целей.

## 2. Этапы прохождения производственной практики:

Таблица 1 ─ Этапы прохождения производственной практики.

| № п/п | Наименование этапа прохождения производственной практики | Содержание этапа | Даты прохождения этапа |
| --- | --- | --- | --- |
|  | Получение индивидуального производственного задания | Получение задания от руководителя практики | 23.06.2022  по 24.06.2022 |
|  | Обработка и анализ полученной информации | Выделение отдельных задач; планирование последовательности и времени их выполнения; определение потребности в источниках информации, необходимых для качественного выполнения задания | 29.06.2022 |
|  | Выполнение практического задания | Создание сайта и разработка его дизайна, написание функци сайта, реализация сбора данных с сайтов (парсинга), создание серверной части веб-приложения | 29.06.2022  по 15.07.2022 |
|  | Анализ итогов практики, подготовка отчета | Подготовка и оформление отчета по производственной практике | 15.07.2022 |
|  | Аттестация | Защита итогов производственной практики | 19.07.2022 |

## 3. Задание для прохождения производственной практики:

**ОБЪЕКТ:** В файле Object.json находится объект с объектами следующего формата: [URL, name, sheet], где:

* URL — ссылка на страницу каталога товаров;
* name — название категории товаров,
* sheet — номер страницы, с которой нужно собирать данные.

**ЦЕЛИ:** Необходимо пройтись по всем URL сайта (и определённым страницам, в зависимости от параметра sheet) “Леруа Мерлен”, сохранив цену товара, URL товара и артикул.

**РЕЗУЛЬТАТ ВЫПОЛНЕНИЯ РАБОТЫ**:

В результате выполнения работы были собраны следующие данные: url изображения товара, url самого товара, наименование товара, цена каждого товара для каждого url страницы (с нужной страницы под номером sheet).

Все полученные данные сохраняются в файле ParsedData.json.

## 4. Подготовка к выполнению задания:

Для выполнения поставленной задачи потребовалось: знание языков HTML, CSS и Javascript, программная платформа Node.js.

Были использованы следующие библиотеки (модули) Node.js: body-parser, express, puppeteer-extra, puppeteer-extra-plugin-stealth.

## 5. Описание выполнения задания по производственной практике:

В коде присутствуют комментарии, объясняющие предназначение и работу функций и большинства частей кода.

Дерево директорий и файлов представленно на рисунке 1.

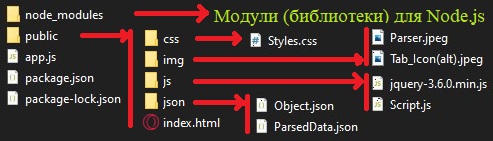


Рисунок 1 ─ Дерево директорий и файлов.

Предназначение основных файлов:

* **app.js** ─ в нем реализован сбор данных с сайтов (парсинг) с помощь библиотеки puppeteer-extra. Библиотека fs используется для работы с файлами (чтение и запись). Библиотека puppeteer-extra-plugin-stealth необходима для доступа к сайту как пользователь, обходя блокировки “подозрительных активностей”. Также реализована серверная часть программы с помощью модуля express.js.
* **index.html** и **Styles.css** ─ HTML код страницы сайта и стили.
* **Object.json** и **ParsedData.json**. В файле Object.json хранится массив объектов, представленный на рисунке 2. Каждый объект имеет 3 параметра: url (ссылка на страницу каталога), name (название каталога) и sheet (номер страницы, с которой нужно собрать данные). В файле ParsedData.json будут храниться все данные, собранные в результате парсинга. Упрощённая структура массива полученных данных представлена на рисунке 3.
* **Parser.jpeg** и **Tab\_Icon(alt).jpeg**. Первое изображение ─ это логотип, который отображается в поле заголовка сайта, а второе ─ используется в качестве изображения вкладки страницы у сайта.
* **jquery-3.6.0.min.js** ─ файл JQuery.
* **Script.js**. Здесь реализованы: прокручивание страницы наверх при нажатии на кнопку в левом нижнем углу страницы, код для анимированного заднего фона страницы и функция добавления к серверу файла ParsedData.json (и правильное размещение этих данных на сайте).



Рисунок 2 ─ Json файл с исходными данными.

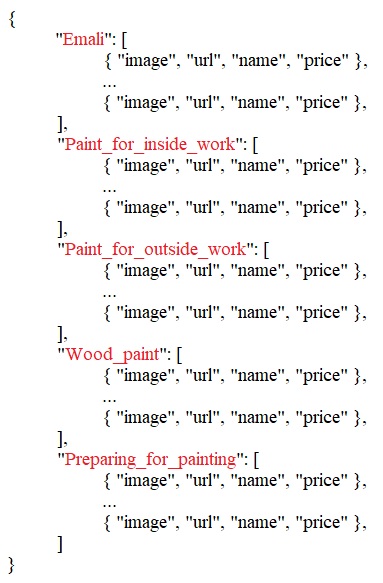


Рисунок 3 ─ Содержание json файла с полученными данными (упрощённое).

Что находится в файле ParsedData.json (рисунок 3):

* "Emali", "Paint\_for\_inside\_work","Paint\_for\_outside\_work", "Wood\_paint" и "Preparing\_for\_painting" ─ названия категорий товаров, представляют собой массивы объектов.
* Каждый объект имеет 4 параметра: image (ссылка на изображение товара), url (ссылка на сам товар), name (наименование товара) и price (стоимость товара).

Рассмотрим содержание файла **app.js**. В ней есть 3 основные функции: **fetchProductList, pageEvaluate** и **startServer**. В первой функции запускается модуль puppeteer и осуществляется переход на нужные страницы сайта Леруа Мерлен, во второй ─ происходит сбор необходимых данных с каждой страницы, в третьей ─ запуск сервера для нашего сайта.

С чего начинается работа программы? Сначала программа считывает файл Object.json, в котором находятся исходные данные, содержимое файла заносится в переменную Myobj. Потом вызывается ассинхронная функция fetchProductList, в которую передаётся значение Myobj.

В функции **fetchProductList** после запуска puppeteer и создания пяти переменных page для страниц сайта Леруа Мерлен (строки 19-33) осуществляется переход по ссылкам, указанных в Object.json (строки 45-49). Когда каждая страница открывается, то программа ждёт, пока будет загружено поле с товарами (строки 58-62). Дальше создаём объект placeHolder, в который будут записываться полученные в результате парсинга данные, и вызываем 5 раз вызываем функцию pageEvaluate (строки 69-74). В эту функцию передаются 3 аргумента: страница, номер этой страницы и объект placeHolder.

В функции **pageEvaluate** для каждой страницы сайта (параметр currentPage) используется метод “.evaluate” (строки 83-123). Внутри него происходит подсчёт количества товаров на текущей странице, потом циклом for пробегаемся по каждому товару, сохраняя следующие данные в объект product: ссылку на изображение товара, ссылку на сам товар, наименование товара и его стоимость. Потом собираем данные из product в другой объект ─ productsList. Метод evaluate возвращает объект productsList, который запишется в константу result. Потом, согласно номеру текущей страницы, в объект obj\_with\_results будут добавлятся определённые параметры (“Emali”, “Wood\_paint” и т.п.), которым будут присваиваться значения переменной result. Когда будет добавлен последний параметр, то все собранные в obj\_with\_results данные сохранятся в файл ParsedData.json, и вызовется последняя функция ─ startServer.

Функция **startServer** не принимает никаких аргументов. В ней происходит подключение модуля express и запуск сайта в качестве приложение на порту port. С помощью app.get реализовано подключение файла ParsedData.json к серверу для дальнейшего чтения данных из него (строки 190-195). К серверу будут подключаться файлы из папки public в качестве статических файлов.

Функции **startServer**, **fetchProductList** и **pageEvaluate** представлены на рисунках 4, 5 и 6 соответсвенно.

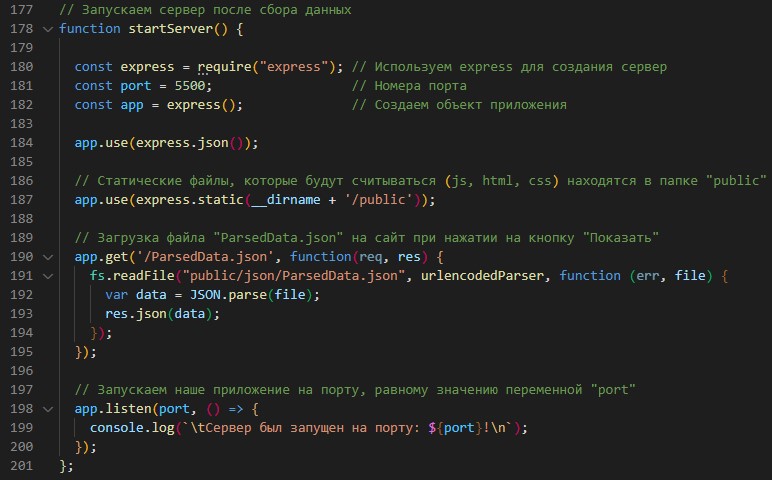


Рисунок 4 ─ Код функции startServer.

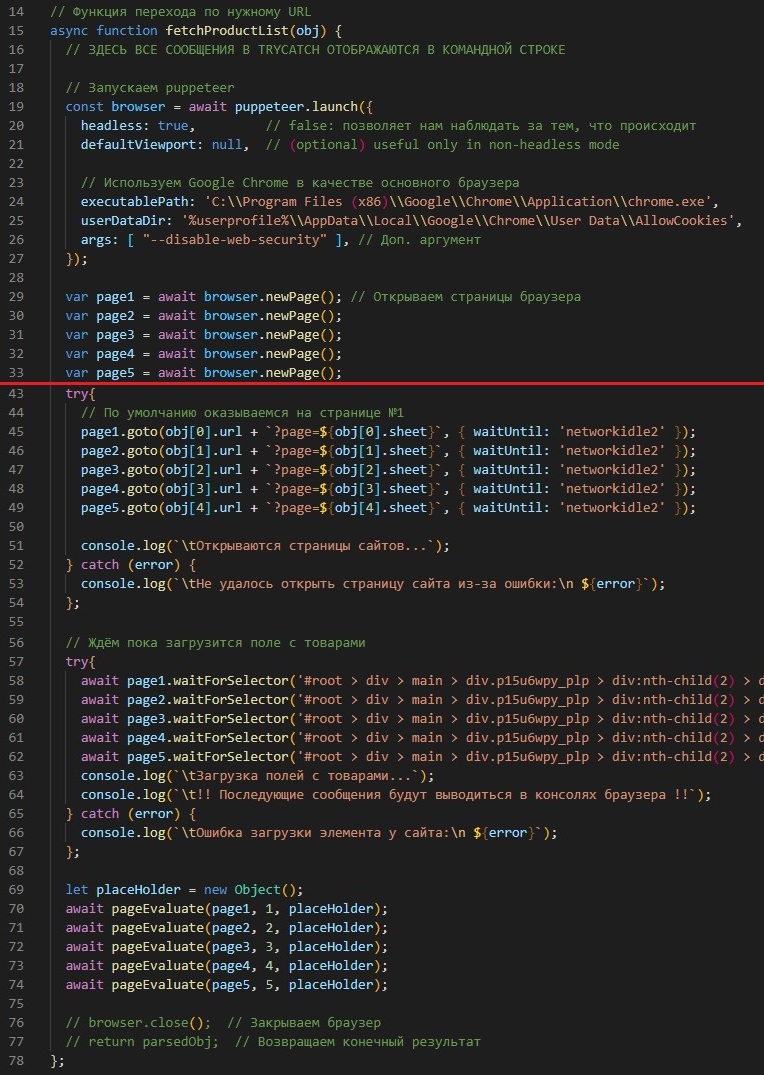


Рисунок 5 ─ Код функции fetchProductList.

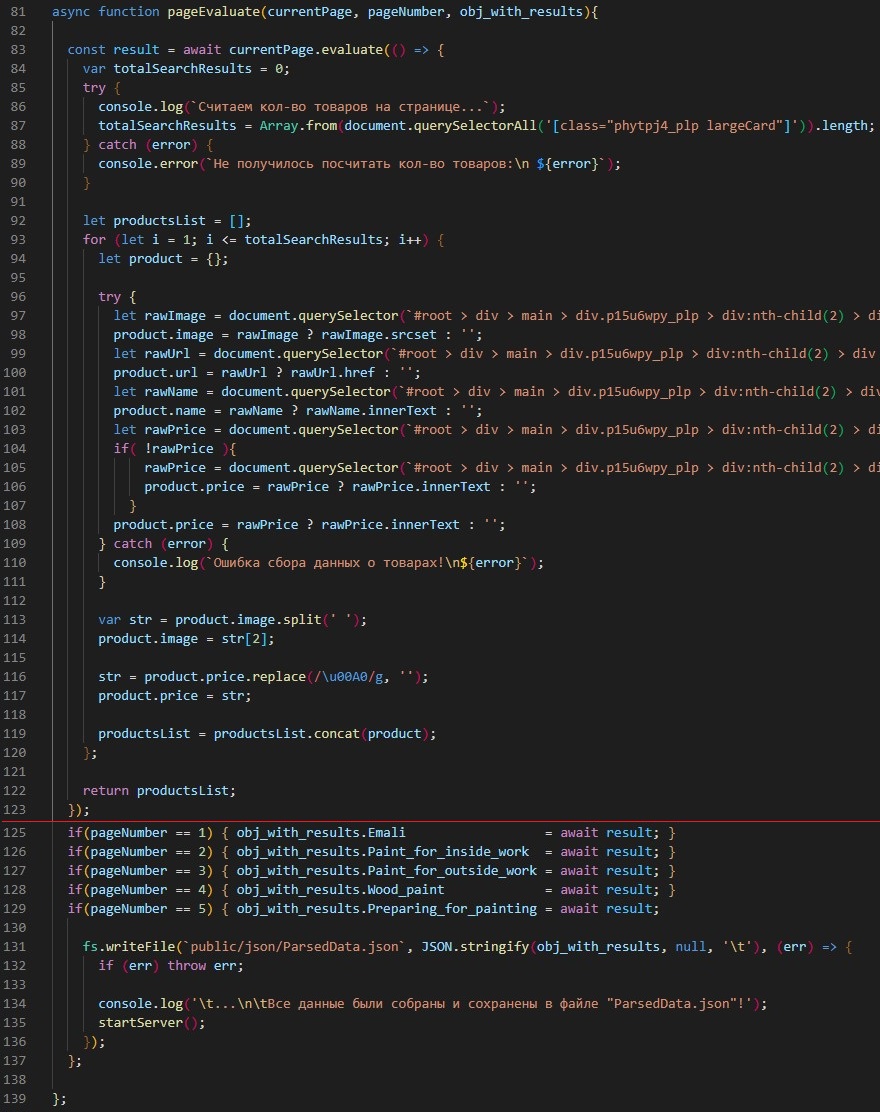


Рисунок 6 ─ Код функции pageEvaluate (без комментариев в коде).

# **ЗАКЛЮЧЕНИЕ**

В ходе прохождения практики была изучена библиотека puppeteer в Node.js, которая позволяет автоматизировать работу с браузером Google Chrome и даёт возможность для создания программ для автоматического сбора данных с веб-сайтов (веб-скраперы), имитирующие действия обычного пользователя

Для меня эта практика оказалось очень продуктивной: она помогла мне развить навыки для работы с такими языками программирования как HTML, CSS и Javascript.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. CTУ 7.5–07–2021 Система менеджмента качества. Общие требования к построению, изложению и оформлению документов учебной деятельности. Красноярск: СФУ, 2021. – 61 с.
2. Puppeteer: парсинг сайтов с JavaScript [Электронный ресурс]. URL: <https://proglib.io/p/puppeteer-parsing-saytov-na-javascript-2020-06-16>.
3. A plugin for puppeteer-extra and playwright-extra to prevent detection [Электронный ресурс]. URL: <https://www.npmjs.com/package/puppeteer-extra-plugin-stealth>.
4. JavaScript, Node, Puppeteer: автоматизация Chrome и веб-скрапинг [Электронный ресурс]. URL: <https://habr.com/ru/company/ruvds/blog/341348/>.
5. Web scraping при помощи Node.js [Электронный ресурс]. URL: <https://habr.com/ru/post/301426/>.