МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«Новосибирский национальный исследовательский государственный университет»

(Новосибирский государственный университет, НГУ)

Структурное подразделение Новосибирского государственного университета –

Высший колледж информатики Университета (ВКИ НГУ)

КАФЕДРА ИНФОРМАТИКИ

**РАЗРАБОТКА ПРОГРАММНОГО СРЕДСТВА ДЛЯ АВТОМАТИЗАЦИИ СБОРА ПУБЛИЧНОЙ ИНФОРМАЦИИ С САЙТА ОПЕРАТОРА ЖЕЛЕЗНОДОРОЖНЫХ КОНТЕЙНЕРНЫХ ПЕРЕВОЗОК И ИНТЕГРАЦИИ С СИСТЕМОЙ ПОСТПРОЦЕССИНГА РАСПОЗНАННЫХ ДАННЫХ**

Квалификация техник-программист

|  |  |
| --- | --- |
| Руководитель  асс. к. инф. ВКИ НГУ | Никифоров А. А.  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_2022 г. |
| Студент 4 курса  803в2 | Маслов Т. И.  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_2022 г. |
| Нормоконтроль | Литвинова О. В.  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_2022 г. |

Новосибирск

2022

**СОДЕРЖАНИЕ**

[ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ 3](#_Toc93557204)

[ВВЕДЕНИЕ 4](#_Toc93557205)

[1 ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ 5](#_Toc93557206)

[2 ПОСТАНОВКА ЗАДАЧИ 7](#_Toc93557207)

[3 АНАЛОГИ 8](#_Toc93557208)

[4 ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ К ПРОГРАММНОМУ ПРОДУКТУ 9](#_Toc93557209)

[5 НЕФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ К ПРОГРАММНОМУ ПРОДУКТУ 10](#_Toc93557210)

[5.1 Требования к программному обеспечению 10](#_Toc93557211)

[5.2 Требования к аппаратному обеспечению 10](#_Toc93557212)

[5.3 Требования к надёжности 10](#_Toc93557213)

[6 ХАРАКТЕРИСТИКА ВЫБРАННЫХ ПРОГРАММНЫХ СРЕД И СРЕДСТВ 11](#_Toc93557214)

[7 АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ 13](#_Toc93557215)

[7.1 Математическая модель 13](#_Toc93557216)

[7.2 Алгоритм реализации программного средства 13](#_Toc93557217)

[7.3 Схема функционирования программного средства 13](#_Toc93557218)

[8 ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ 15](#_Toc93557219)

[8.1 Описание базы данных 15](#_Toc93557220)

[9 ОТЛАДКА И ТЕСТИРОВАНИЕ ПРОГРАММНОГО СРЕДСТВА 17](#_Toc93557221)

[10 РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ 18](#_Toc93557222)

[ЗАКЛЮЧЕНИЕ 19](#_Toc93557223)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 21](#_Toc93557224)

[ПРИЛОЖЕНИЕ А 22](#_Toc93557225)

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ

Парсинг – автоматизированный сбор информации из источника, содержащего дополнительную или избыточную информацию, а также структурирование полученного содержимого.

Парсер – программное средство, которое выполняет парсинг.

HTML-представление – документ веб-страницы в браузере основанный на стандартизированном языке разметки HTML и соответствующий тому, как веб-страница отображается пользователю.

Капча (CAPTCHA) – компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы: человеком или компьютером.

JSON – это формат, который хранит структурированную информацию и чаще всего используется для передачи данных между сервером и клиентом.

Cross-origin resource sharing (CORS) – технология современных браузеров, которая позволяет предоставить веб-страницам доступ к ресурсам другого домена.

****ВВЕДЕНИЕ****

Основная проблема современного Интернета – избыток информации, которую человек не в состоянии систематизировать вручную. Для автоматического сбора и анализа большого объёма информации существует программное обеспечение – парсер. Он собирает требуемую пользователем информацию и выводит нужные результаты. Парсер может использоваться для:

1. Анализа ценовой политики. Чтобы определить среднюю стоимость товаров на рынке, используя информацию от конкурентов. Если таких позиций сотни или тысячи штук, собрать вручную такое количество информации крайне затруднительно.
2. Отслеживание изменений. Можно на регулярной основе следить за средней ценой на рынке среди конкурентов.
3. Исправление ошибок собственного сайта. С помощью парсера можно найти среди тысячи товаров дубли, несуществующие страницы и многое другое.
4. Заполнение карточек товаров. Заполнение информации для каждого из тысячи товара вручную будет проблематично. Парсер может на основе других сайтов собрать описание к товару автоматически.

Данная работа нацелена на создание парсера, который собирает информацию с сайта на основе введённых пользователем данных.

****1 ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ****

Слово «парсинг» с английского языка (to parse) означает «разбор» и применяется для описания того, как из содержимого произвольного формата (например, текста) выделяются машиночитаемые структуры (числа, XML-деревья и так далее). Парсить – собирать и систематизировать информацию с сайтов с помощью специальных программ и скриптов (парсеров).

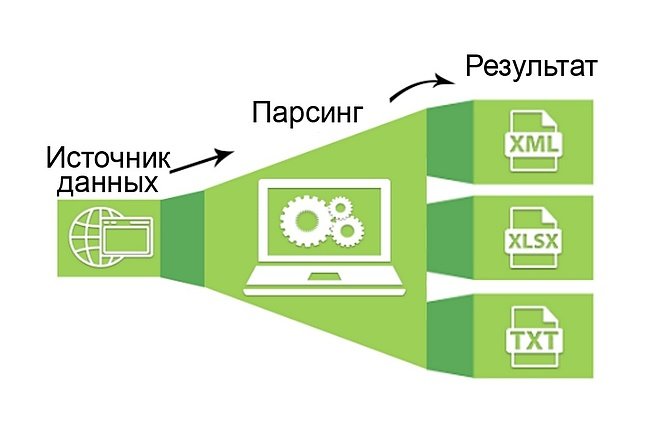


Рисунок 1.1 – Обобщённый процесс парсинга

Специальные программы или сервисы-парсеры «обходят» сайт и собирают данные, которые соответствуют заданному условию.

Простой пример: нужно собрать контакты потенциальных партнеров из определенной ниши. Вы можете это сделать вручную. Надо будет заходить на каждый сайт, искать раздел «Контакты», копировать в отдельную таблицу телефон и т. д. Так на каждую площадку у вас уйдет по пять-семь минут. Но этот процесс можно автоматизировать. Задаете в программе для парсинга условия выборки и через какое-то время получаете готовую таблицу со списком сайтов и телефонов.

Плюсы парсинга очевидны – если сравнивать его с ручным сбором и сортировкой данных:

* данные могут быть получены очень быстро;
* можно задавать десятки параметров для составления выборки;
* в отчете не будет ошибок;
* парсинг можно настроить с определенной периодичностью – например, собирать данные каждый понедельник;
* многие парсеры не только собирают данные, но и советуют, как исправить ошибки на сайте.

****2 ПОСТАНОВКА ЗАДАЧИ****

Целью данной работы является разработка программного средства позволяющий на основе введённых данных собирать информацию с интернет ресурса [https://isales.trcont.com/#/].

Для достижения цели поставленной задачи требуется получить исходный код веб-страницы, вставить данные, введенные пользователем в соответствующие поля, отправить POST запрос, собрать выведенную информацию и структурировать её в требуемом виде для клиента.

****3 АНАЛОГИ****

Существует два вида парсеров:

* облачные – чтобы получить доступ к такому парсеру достаточно перейти на сайт производителя, где будет выведен весь нужный интерфейс;
* локальные – парсер устанавливается непосредственно на компьютер клиента;

Все парсеры для веб-содержимого создаются под конкретный сайт с конкретным сбором информации. Как правило, парсеры создаются под конкретного заказчика и не предполагаются к публичному распространению.

****4 ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ К ПРОГРАММНОМУ ПРОДУКТУ****

Программное обеспечение должно иметь возможность встраивания/интеграции с веб-сервисами третьей стороны.

Программное обеспечение должно загружать с целевого адреса информацию, доступную неопределённому кругу лиц. В частности, программное обеспечение не должно требовать установленного сеанса авторизации/аутентификации.

Программное обеспечение должно преобразовывать введённую информацию в не интерактивное состояние, доступное конечному пользователю (DOM-дерево, HTML-документ, статическая JSON-строка).

Программное обеспечение должно выполнять разбор полученного не интерактивного значения и возвращать целевые данные.

Программное обеспечение должно сериализовать полученные данные документированным образом.

Программное обеспечение должно передавать/экспортировать полученные данные для интеграции со сторонними сервисами.

Ошибки соединения должны быть обработаны документированным образом и не порождать исключительных ситуаций.

****5 НЕФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ К ПРОГРАММНОМУ ПРОДУКТУ****

5.1 Требования к программному обеспечению

* Возможность использовать JavaScript;
* Отсутствие ограничения доступ к целевым сайтам;

5.2 Требования к аппаратному обеспечению

* Наличие открытого доступа в сеть Интернет;
* Оперативная память размером не менее 1 Гб;

5.3 Требования к надёжности

* Программное обеспечение не должно передавать данные третьим сторонам, кроме тех, которые обоснованы в документации;
* Программное обеспечение не должно прекращать работу, столкнувшись с недоступностью целевого сайта, изменением протокола обмена или другой нештатной ситуацией;
* Программное обеспечение не должно транслировать исключительную ситуацию;

****6 ХАРАКТЕРИСТИКА ВЫБРАННЫХ ПРОГРАММНЫХ СРЕД И СРЕДСТВ****

JavaScript очень популярен благодаря своим преимуществам.

* Незаменимость для веб-разработки. Поддержка скриптов всеми популярными браузерами; полная интеграция с вёрсткой страниц (HTML+CSS) и серверной частью.
* Скорость работы и производительность. Javascript позволяет частично обрабатывать веб-страницы на компьютерах пользователя без запросов к серверу. Это экономит время и трафик, снижает нагрузку на сервер.

Как любой язык программирования, Javascript имеет некоторые недостатки или ограничения.

* Нет возможности чтения и загрузки файлов. Это ограничение функциональности на стороне клиента. Главная причина – соображения безопасности.
* Нестрогая типизация и вольная трактовка. Язык игнорирует явные нестыковки. Имеет место разная интерпретация данных. Нет возможности раннего выявления ошибок. Все недочёты выявляются уже на этапе работы.
* Нет поддержки удалённого доступа. Поэтому язык нельзя использовать для сетевых приложений. За это Javascript даже не считают полноценным языком программирования.
* Доступность для злоумышленников. В свободный скриптовый язык проще всего встроить фрагмент вредоносного кода, который может навредить пользователю.

Visual Studio Code – редактор исходного кода, разработанный Microsoft для Windows, Linux и macOS. Позиционируется как «лёгкий» редактор кода для кроссплатформенной разработки веб- и облачных приложений. Благодаря поддержке пользовательских расширений, он очень гибок при настройке и удобен в дальнейшем использовании.

Mozilla Firefox – бесплатный браузер на движке Quantum, разработкой и распространением которого занимается Mozilla Corporation. Имеет удобные инструменты для анализа сайта и запросов, связанных с ним.

****7**** АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ

{В этом разделе необходимо описать используемые математические алгоритмы решения задачи (если они есть), методы реализации ПС или ПТС (со схемами) и исследовательскую часть, если она необходима для решения проблемы. Вся информация может быть описана как в одном разделе, так и разбита, например, на нижеприведенные подразделы.}

7.1 Математическая модель

{В этом подразделе необходимо описать используемые математические алгоритмы решения задачи с формулами и со ссылками на источники. Правила оформления формул и ссылок приводятся в метод.пособии Родионовой О.К.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

7.2 Алгоритм реализации программного средства

{Для описания алгоритма реализации ПС или ПТС могут использоваться два способа описания алгоритмов:

- словесный: в виде последовательности шагов по реализации с описанием классов, интерфейсов, процедур (можно с программным кодом, реализующим наиболее интересные алгоритмы). Если это техническая часть, то помимо словесного описания могут быть приведены схемы технических устройств;

- графический: в виде блок-схем обязательно с пояснениями.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

7.3 Схема функционирования программного средства

{В этом разделе может быть полностью или частично описана структура ПС или ПТС, т. е.:

- из каких функциональных блоков (файлов, модулей, процедур, функ-ций, классов) состоит ПС или ПТС;

- приведено описание каждого блока с его названием и назначением;

- приведена графическая схема взаимосвязи этих блоков.

Для описания схемы функционирования программного средства можно использовать диаграммы UML.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

****8 ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ****

{В данном разделе описывается структура входных, выходных и про-межуточных данных. Например, входные данные поступают на вход программы в виде файла, значит, структура файла должна быть полностью описана. Аналогичное требование выполняется для промежуточных и выходных данных. Например, для сайтов входной информацией могут быть текстовые файлы (указать в каком формате), графическая информация (указать формат) и т. п. Выходная информация – это HTML, PHP и т. п. страницы, видимые в окне браузеров (указать, каких).}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

8.1 Описание базы данных

{В этом разделе необходимо описать структуру базы данных, если таковая имеется в проекте. Если приложение не использует базу данных, этот пункт опускается.

При описании структуры базы данных описываются все таблицы в виде:

- имя таблицы и её назначение, т.е. для хранения какой информации предназначена данная таблица;

- описание всех полей таблицы с указанием типа, назначения, первичных и внешних ключей.

Пример описания структуры и концептуальной схемы базы данных приведен в приложении Г в метод.пособии Родионовой О.К.

Для объектной базы данных приводится описание структуры данных всех классов, аналогично тому, как это сделано для таблиц, добавляется только описание методов.

Если в выпускной квалификационной работе разработаны и созданы инструменты для работы с БД, их можно описать в данном разделе.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

****9 ОТЛАДКА И ТЕСТИРОВАНИЕ ПРОГРАММНОГО СРЕДСТВА****

{В этом разделе необходимо указать степень отлаженности программного или технического средства. Например, указать, что программное средство полностью отлажено и сдано в эксплуатацию (например, опытную). Можно указать, прошло оно альфа или бета-тестирование, т. е., протестировано только разработчиком или "посторонними" коллегами. Также можно указать на каких данных, каким образом проводилось тестирование ПС или ПТС. При необходимости можно привести тест на правильность функционирования ПС или ПТС. Тесты специально разрабатываются для проверки всевозможных ситуаций работы разрабатываемого ПС или ПТС, документируются и прилагаются результаты тестирования. По результатам тестирования ПС или ПТС может быть проведена техническая, алгоритмическая или программная оптимизация.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

****10 РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ****

{В этот раздел могут быть включены следующие пункты:

- порядок установки и настройки программного средства на компьютер пользователя или первичная настройка технического средства;

- действия пользователя в случае сбоя ПС или ПТС;

- последовательность действий пользователя для решения своих основных задач при работе с программным или техническим средством.

Следует привести образы экранов (ScreenShots) с пояснениями для лучшего восприятия раздела. Приветствуется создание помощи по установке и эксплуатации ПС или ПТС.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

ЗАКЛЮЧЕНИЕ

{Заключение – последовательное, логически стройное изложение полученных итогов и их соотношение с общей целью и конкретными задачами, поставленными и сформулированными во введении.

Заключение должно содержать краткие выводы по результатам исследования, отражающим новизну и практическую значимость работы, предложения по исполь-зованию ее результатов, технико-экономической эффективности.

В заключении суммируют теоретические и практические выводы, а также те предложения, к которым автор пришел в результате проведенного исследования. Выводы формулируются по пунктам так, как они должны быть оглашены в конце доклада на защите.

Заключение отражает оценку работы и включает рекомендации по практиче-скому использованию ее результатов.

- В этом разделе необходимо указать решена задача полностью или ча-стично:

- подвести итоги проделанной работы – что сделано для решения постав-ленной задачи, это может быть:

а) разработанное и реализованное ПС или ПТС;

б) научно-исследовательская работа;

в) список решенных модельных или реальных задач;

г) разработаны новые математические алгоритмы;

д) предложен новый подход к решению подобных задач;

е) предложен новый метод реализации ПС или ПТС;

ж) разработана технология решения подобных задач.

- оценить практическую значимость работы, будет ли она иметь практи-ческое применение;

- отметить возможные точки роста (развития) ПС или ПТС. Например:

а) добавить новые функции (блоки, режимы работы);

б) распространить на новый класс задач, другой тип данных.

- привести сравнительные характеристики (кратко) проделанной работы с существующими аналогами, если таковые имеются;

- перечислить виды проделанной работы и полученные результаты;

- привести объем созданного программного средства (в любых единицах, например, в килобайтах, строках кода, классах, процедурах, функциях ит.д.).

- отразить апробацию работы: выступления на семинарах, публикации, выступления на конференциях и конкурсах (полученные дипломы, грамоты и т.п.), акты о внедрении, реально действующий сайт (адрес) и т.д.}

Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст Ваш текст

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

{Сведения об источниках можно располагать в порядке:

- появления ссылок на источники в тексте документа.

Источники необходимо нумеровать арабскими цифрами без точки и печатать с абзацного отступа.

Ссылки на использованные источники следует приводить в квадратных скобках в тексте, в конце предложения.

Примеры библиографического описания источников (более подробная инф. в метод.пособии Родионовой О.К.):}

1. Численное исследование процессов переноса и трансформации газовых и аэрозольных примесей в шлейфе выбросов Норильского промышленного района / В.Ф. Рапута, В. Симоненков, Б.Д. Белан, Т.В. Ярославцева / «Оптика атмосферы и океана», 31, № 6, 2018 –– C. 438-439.
2. Вострикова Л.Г. Финансовое право [Электронный ресурс] : учебник / Л.Г. Вострикова. – Электрон. текстовые дан. – М.: Равновесие: Юстинформ, 2005. – 1 электрон. опт. диск (СD). – Загл. с контейнера.
3. Исследовано в России[Электронный ресурс]: многопредмет. науч. журн. / Моск. физ.-техн. ин-т.–Электрон. журн. – Долгопрудный : МФТИ, 1998.– URL : http://zhurnal.mipt.rssi.ru, свободный. –Яз. рус. – (Дата обращ. ……..)
4. Мозолин В. П. Модернизация права собственности в экономическом измере-нии [Электронный ресурс] / В. П. Мозолин. – Электрон. ст. – М., 2011. – URL : http://www.norma-verlag.com/journal/2011/1, свободный. – Яз. рус. – Аналог печат. изд. (Журнал российского права. – 2011. – № 1). – (Дата обращ. 23.06.2011).
5. Российская государственная библиотека [Электронный ресурс] / Центр ин-форм. технологий РГБ; ред. Т.В. Власенко; Web-мастер Н.В. Козлова.–Электрон. дан.– М.: Рос. гос. б-ка, 1997. – URL : http//www.rsl.ru, свобод-ный.–Яз. рус., англ. – (Дата обращ. ………).

ПРИЛОЖЕНИЕ А

{В приложения рекомендуется включать материалы, связанные с выполненным дипломным проектом, которые по каким-либо причинам не могут быть включены в основную часть.

В приложения могут быть включены:

- тексты программ;

- иллюстрации вспомогательного характера;

- таблицы вспомогательных цифровых данных;

- протоколы испытаний;

- описание аппаратуры и приборов, применяемых при проведении экспериментов, измерений и испытаний;

- тесты.}

Текст программы

Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст. Ваш текст.