**МИНОБРНАУКИ РОССИИ**

**федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова»**

**(БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | |  | | | | | | | | | | |
| Факультет | O | | Заведующий кафедрой | | | | | |  | | O7 | | |
|  | индекс | |  | | | | | |  | | индекс кафедры | | |
| Выпускающая кафедра | O7 | | Семенова Е.Г | | | |  |  | | | | | |
|  | индекс | | Фамилия И.О. | | | |  | подпись | | | | | | |
| Группа | И595 | | «\_\_\_\_\_» | |  | | | | | | | 20\_\_\_ г. | |
|  | | номер группы | |  | |  | |  | | | |  | | |

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**ПО ПРОГРАММЕ БАКАЛАВРИАТА**

|  |  |
| --- | --- |
| Васильева Юрий Антоновича | |
| Фамилия, имя, отчество обучающегося в родительном падеже | |
| **На тему** | Разработка приложения для анализа вакансий и резюме с сайта |
| hh.ru | |
|  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Направление подготовки** | | 09.03.04 |  | Программная инженерия |
|  | | индекс направления |  | полное наименование направления |
|  | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Профиль** | | Разработка программно-информационных систем | | |
|  | | | наименование профиля |
|  | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Консультант:** | | |  | | | | | | | | |  | | **Руководитель:** | | | |  | | | | | | | | | |
| *если утвержден приказом* | | | подпись | | | | | | | | |  | |  | | | | подпись | | | | | | | | | |
|  | | | |  |  | | | | | | | |  | |  | | | | |  | |  | | | | | | |
| ученая степень, ученое звание | | | | | Фамилия ИО | | | | | | | |  | | ученая степень, ученое звание | | | | | | Фамилия ИО | | | | | | | |
| « » |  |  | | | | |  | 20 |  | г. |  | |  | | « » |  |  | | | | | |  | 20 |  | г. |  | |
|  | | | | | |  | | | | | | |  | | **Обучающийся:** | | | |  | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | |  |  | | |  |  | | |  |  | | | | | |
|  | | | |  | | |  | подпись | | |  | Фамилия ИО | | | | | |
|  |  |  | | |  |  |  | « » |  |  | | |  |  | 20 |  | г. |

САНКТ-ПЕТЕРБУРГ

2023 г.

# СОДЕРЖАНИЕ

[ВВЕДЕНИЕ](#_bookmark0) 3

1. [Анализ предметной области](#_bookmark4) 5
   1. [Постановка задачи](#_bookmark5) 5
   2. [Описание предметной области](#_bookmark6) 6
   3. [Анализ систем автоматического тестирования](#_bookmark8) 8
   4. [Сравнительный анализ аналогов](#_bookmark9) 9
2. [Составление технического задания на разработку программного продукта](#_bookmark2) 12

[Приложение А. Техническое задание](#_bookmark11) 13

**ВВЕДЕНИЕ**

За последнее десятилетие данные стали ресурсом для развития бизнеса, а Интернет — их основным источником благодаря пяти миллиардам пользователей, формирующим миллиарды фрагментов данных каждую секунду. Анализ данных Всемирной паутины может помочь компаниям выявлять скрытые закономерности, позволяющие им добиваться выполнения своих целей. Однако сбор большого объема данных — непростая задача для новой компании на рынке.

На любом предприятии один из важнейших ресурсов – человеческий. Зачастую новые предприятия или малый бизнес, активно развивающие на рынке труда, обращаются за поиском новых сотрудников в кадровые агентства, которые и занимаются тем самым сбором большого объема информации о вакансиях и резюме.

Но сегодня тяжело представить кадровое агентство, работающее без внедрения программного обеспечения для автоматического сбора данных вакансий и резюме, их анализа и предоставления точного и готового отчета для своих клиентов.

Парсинг веб-сайтов по поиску работы или сотрудников позволяет компаниям автоматизировать процессы сбора данных во Всемирной паутине, используя ботов или автоматические скрипты, называемые «обходчиками» веб-страниц, автоматическими сборщиками данных или веб-сборщиками (web crawlers) [1]-[3].

Подытожив, можно сказать, что актуальность темы обосновывается в необходимости отойти от ручного поиска вакансий или резюме и внедрение автоматического ПО, чтобы избежать некорректного заполнения данных и проверки достоверности найденной информации. Актуальность темы дипломной работы связана со значительным распространением такого явления и заключается в необходимости разработки способа ускорения поиска в рассматриваемой области.

Объект исследования: процесс автоматического сбора данных.

Предмет исследования: системы автоматического сбора данных.

Целью выпускной квалификационной работы является разработка программного обеспечения для автоматического сбора данных с сайта hh.ru вакансий и резюме и их анализа.

Для достижения поставленной цели необходимо было решить следующие задачи:

* изучить принципы сбора данных;
* изучить существующие системы автоматического сбора данных;
* разработать систему требований;
* выполнить проектирование;
* разработать пользовательский интерфейс;
* выполнить программную реализацию системы;
* протестировать работоспособность системы.

ВКР состоит из введения, трех разделов и заключения.

В первом разделе приводятся постановка задачи, изучение принципов и систем автоматического сбора данных, разработка системы требований.

Во втором разделе приводится описание процесса проектирования системы сбора данных.

Третий раздел содержит обоснование выбора средств разработки, описание реализации программного продукта и сведения о результатах сбора данных.

**1 Анализ предметной области**

В этом разделе рассматриваются предметная область и существующие системы автоматического сбора данных.

* + - 1. **Постановка задачи**

Требуется разработать приложение автоматического сбора данных вакансий и резюме.

В первую очередь, в приложении должен выполняться поиск по вакансиям с запрашиваемыми фильтрами данных, введенные пользователем, а именно:

* название вакансии;
* заработная плата;
* график работы;
* регион;
* опыт работы;
* образование;
* количество вакансий, которые нужно найти.

Приложение должно выполнять поиск по резюме с запрашиваемыми фильтрами данных, введенные пользователем, а именно:

* специализация;
* регион;
* заработанная плата;
* требуемы опыт работы;
* образование;
* график работы;
* тип занятости;
* знание языков;
* количество резюме, которые нужно найти.

Пользователю важно иметь возможность редактирование уже введенных данных в целях корректировки значений и вывода последующей желаемой информации.

Пользователь должен иметь управление приложением через телеграм бота. Необходимо реализовать возможность задавания параметров, вывода и сохранение результатов, не имея возможности пользоваться интерфейсом приложения на ПК.

Пользователю необходимо предоставить удобный интерфейс, позволяющий легко ориентироваться в приложение. Желательны минимальное количество одновременных элементов интерфейса на экране и неяркие цвета, использование яркого цвета на активном объекте. Это поможет сосредоточить внимание. Контекстные подсказки необходимы для предотвращения ошибок в заполнении форм пользователем. Необходимо создать в макете интерфейса для каждого критерия найденной информации с веб-ресурса блок для вывода данной информации. Следующие блоки вывода информации: «Название вакансии», «Требуемый опыт работы», Заработанная плата», «Занятость», «Адрес», «Ссылка на вакансию». Также необходимо создать меню, в котором будут реализованы функции сохранения информации в CSV файл или текстовый файл, сохранение логов в отдельный текстовый файл. Создание отдельного блока с кнопками: «Запуск», «Стоп», «Пауза», «Снять с паузы», «Просмотр логов».

Пользователь важно также иметь возможность просмотра промежуточных результатов во время работы приложения.

Приложение должно сохранять информацию в CSV-таблицу или текстовый файл.

* + - 1. **Описание предметной области**

Парсер — это программа для сбора и систематизации информации, размещенной на различных сайтах. Источником данных может служить текстовое наполнение, HTML-код сайта, заголовки, пункты меню, базы данных и другие элементы. Процесс сбора информации называется парсинг (parsing).[4]

Объектом парсинга может быть любая грамматически структурированная система: информация, закодированная естественным языком, языком программирования, математическими выражениями и т.д. Например, если исходный массив данных представляет собой HTML-страницу, парсер может вынести из кода информацию и перевести ее в текст, понятный для человека. Или конвертировать в [JSON](https://blog.skillfactory.ru/glossary/json/) — формат для приложений и скриптов[5]

При составлении запроса для парсера решаются следующие вопросы:

* какую информацию необходимо собирать;
* какое количество необходимой информации нужно собрать;
* время проведения парсинга.

При разработке сайтов с вакансиями и резюме используется язык HTML, типовые элементы блоков. Поэтому стандартные парсеры данных используют типовые теги HTML, которые имеются на страницах о вакансиях и резюме:

* тег div. Этот блочный элемент дает возможность выделить раздел, в котором есть визуальный контент;
* тег a. Отображение ссылок на продукцию выбранной товарной категории.
* тег h1 и др. Отображает заголовки различных уровней. Заголовками могут быть названия вакансий, резюме, адреса и другие блоки;
* тег p. Служит для отображения текстового абзаца — описаний вакансии или резюме.
* тег table. Отображает таблицы;
* тег ul. Обозначает маркированный список (перечень навыков, языков и т.п.);
* тег img. Отображает фотографии, другие изображения [6]-[7].

Используя стандартизированные настройки, можно выбрать именно те блоки информации, которые нужны для анализа или других целей.

Применение парсинговых утилит позволяет быстро собирать и систематизировать большие массивы данных.

* + - 1. **Анализ систем автоматического сбора данных**

Система автоматического сбора данных представляет собой программу, содержащую в себе инструменты для проведения процесса сбора данных. Появлению таких систем поспособствовала анализ динамики изменений, сбор ссылок и составление клиентской базы, и сегодня они используются для ускорения процесса сбора данных.

Функции систем автоматического сбора данных обычно следующие:

* получение HTML-разверстки [8];
* изменение параметрами вводимых данных;
* авторизация в качестве пользователя;
* просмотр результатов автоматического сбора данных.

Авторизация по логину и паролю требуется для идентификации пользователя.

Процесс автоматического сбора данных схож для всех систем. Описать его можно следующим образом.

* сначала пользователь указывает вводные данные для парсинга;
* затем указывает список страниц или ресурсов, на которых нужно осуществить поиск;
* после этого программа в автоматическом режиме проводит глубокий анализ найденного контента и систематизирует его;
* в итоге пользователь получает отчет в заранее выбранном формате.

Автоматический сбор данных имеет преимущества, такие как:

* экономия времени [9];
* исключение влияния человеческого фактора на результат [10].

Среди самых существенных недостатков автоматического сбора данных обычно выделяют следующее:

* наличие защиты от данного способа индексирования информации у определенных IP-адресов и вытекающая из этого невозможность сбора всех нужных данных [11];
* возможные перерывы в работе из-за многочисленности пользователей на сайте. Это происходит вследствие того, что парсер не может обновить веб-страницу для оптимальной загрузки.
  + - 1. **Сравнительный анализ аналогов**

При выборе приложения парсера вакансий и резюме наиболее разумно ориентироваться на такие ее показатели как возможности покупки данного приложения, цена, удобство, время выполнения работы приложения.

В настоящее время существует очень малое количество автоматизированных информационных систем в свободном доступе, осуществляющих решение задачи автоматизации сбора данных по вакансиям и резюме.

Прежде всего рассмотрим работу приложений автоматического сбора данных информации по вакансиям и резюме на примере двух готовых решений: сервис Datacol и ParserJobSites.

Программное обеспечение «Datacol», в данной системе реализованы следующие возможности [12]:

* может производить поиск информации по вакансиям и резюме, используя ссылку на каталог по интересующей вакансии или резюме;
* приложение платное, необходимо приобретение подписки для активации полного доступа к функциям приложения в размере 633 рубля в месяц, но имеет демо-версию, где сохраняет 25 результатов парсинга;
* имеет возможности добавление прокси-сервера или VPN;
* есть система авторизации для клиентов, купивших данный программный продукт;
* имеет возможность остановки, возобновления и досрочного завершения работы программы;
* экспорт данных производится в различные форматы: CSV-файл, текстовый файл, базу данных MySQL и другие;
* есть возможность изменения собираемых данных.

Также похожим вариантом является приложение «ParserJobSites», в данной системе реализованы следующие возможности [13]:

* может производить поиск только по вакансиям;
* парсинг производится по вакансиям актуальным в поиске на данный момент, нет возможности ввести интересующую вакансию;
* данная программа платная, стоимость составляет 2000 рублей, также имеет демо-версию;
* не имеет возможности добавление прокси-сервера или VPN;
* имеет возможность остановки, возобновления и досрочного завершения работы программы;
* экспорт данных производится в CSV-файл.

Для сопоставления параметров рассмотренных аналогов приведено в таблице 1.

Таблица 1 - Сравнительный анализ программных продуктов для автоматического сбора данных

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование критерия | Datacol | ParserJobSites |
| 1 | Удобство использования | +/- | +/- |
| 1.1 | Удобство изучения | + | + |
| 1.2 | Понятность | +/- | +/- |
| 1.3 | Удобство установки | +/- | +/- |
| 2 | Функциональность | + | - |
| 3 | Надежность | + | - |
| 4 | Эффективность | + | - |

Таким образом, рассмотренные аналоги не подходят для решения поставленной задачи, так как нет необходимости в большинстве функций данных систем, а также основной целевой аудиторией являются пользователи, которые хотят сократить количество расходов.

Поэтому было принято решение разработать собственный программный продукт, в котором присутствует фактор использования дистанционного режима управления в случае отсутствия возможности использования приложения на персональном компьютере, где задание нужных параметров, выгрузка и сохранение информации происходит непосредственно в приложение Telegram через бота. Появляется вариант использования приложения на различных устройствах: мобильное устройство, персональный компьютер, веб-версия приложения Telegram [14]-[15].

# Составление технического задания на разработку программного продукта

В процессе составления технического задания на разработку программного продукта были сформулированы следующие группы требований:

* требования к составу выполняемых функций;
* требования к надежности;
* требования к составу и параметрам технических средств;
* требования к информационной и программной совместимости;
* требования к маркировке и упаковке;
* требования к программной документации.

Полный текст технического задания представлен в приложении А.

Приложение А. Техническое задание

1. **Введение**
   1. Наименование

Система автоматизированного сбора информации и фильтрация данных, далее «Система».

* 1. Краткая характеристика

Система предназначена для сбора информации о вакансиях и резюме с сайта hh.ru.

1. **Основания для разработки технического задания на ВКР.**

Основанием для разработки является техническое задание на выпускную квалифицированную работу.

1. **Назначение разработки**
   1. Функциональное назначение

Система будет использоваться внутри организации следующими лицами: сотрудниками кадровых агенств.

Пользователь имеет возможность задать фильтры поиска для нахождения нужной информации по интересующей вакансии или резюме.

Пользователь задает формат выводимой информации

* 1. Эксплуатационное назначение

Система работает на сервере предприятия с выходом в интернет.

1. **Требования к программе или программному изделию**
   1. **Требования к функциональным характеристикам**
      1. Требования к составу выполняемых функций

* проверка соединения сервера с веб-ресурсом;
* редактирование и удаление вводимых данных, в случае изменения параметров поиска;
* взаимодействие с приложением через приложение Telegram с помощью Telegram-бота (реализация основных функций приложений через Telegram-бота);
* функция вывода результата по заданным параметрам;
* отображение статистики;
* функция сохранения в файл;
* сохранение промежуточных результатов;
* функция создания отчета о тестировании.
  + 1. Организация входных и выходных данных

Ввод информации происходит с клавиатуры в блоке вводимых параметров значений.

Информация, полученная в результате работы программы, хранится в CSV-файле.

* 1. **Требования к надежности**

Система должна работать исправно при условии отсутствия неполадок в локальной сети предприятия.

Система должна быть проста или иметь встроенные подсказки и инструменты помощи во избежание ошибок в работе Системы, которые могут быть вызваны работой тестируемых или администратора.

* 1. **Условия эксплуатации**

Система должна запускаться на локальном сервере предприятия.

Необходим один сотрудник на роль администратора, от которого требуется хотя бы базовый уровень компьютерной грамотности. Предпочтительнее, чтобы это был человек, хорошо знакомый с компьютером, с базовым пониманием компьютера.

* 1. **Требования к составу и параметрам технических средств**

Требования к серверу:

* + процессор с тактовой частотой не менее 1.6 ГГц;
  + оперативная память не менее 8 Гб;
  + жесткий диск объемом не менее 40 Гб;
  + подключение к интернету со скоростью передачи данных не менее 100 Мбит/с.

Требования к пользовательскому компьютеру:

* + процессор с тактовой частотой не менее 1 ГГц;
  + оперативная память не менее 2 Гб;
  + видеокарта;
  + монитор;
  + мышь;
  + клавиатура.
  1. **Требования к информационной и программной совместимости**

Сообщения между сервером и клиентами происходят в локальной сети,

при этом используется протокол HTTP.

Для работы системы на сервере требуется наличие: OC Linux Ubuntu/ Windows 10, Chromedriver последней версии, python 3.7.9.

* 1. **Требования к маркировке и упаковке**

Система передается посредством сети Internet при помощи GitHub.

1. **Требования к программной документации**

Требуется разработать руководство администратора для ускоренного освоения программы и предотвращения некорректной работы.

1. Стадии и этапы разработки

Таблица 1 – Стадии, этапы и их содержание

|  |  |  |
| --- | --- | --- |
| Стадии разработки | Этапы работ | Содержание работ |
| 1. Техническое задание | Обоснование необходимости разработки программы | Постановка задачи.  Обоснование необходимости проведения научно-исследовательских работ. |
| Научно- исследовательские работы | Определение структуры входных и выходных данных.  Определение функциональных характеристик.  Предварительный выбор методов решения задач.  Определение требований к техническим средствам. |
| Разработка и утверждение технического задания | Определение требований к программе. Определение стадий, этапов разработки программы и документации на неё.  Выбор языков программирования. Определение необходимости проведения научно-исследовательских работ на последующих стадиях.  Согласование и утверждение технического задания. |
| 2. Эскизный проект | Разработка эскизного проекта | Предварительная разработка структуры входных и выходных данных.  Разработка пользовательского интерфейса.  Уточнение методов решения задачи. Разработка общего описания алгоритма решения задачи |
| 3. Технический проект | Разработка технического проекта | Уточнение структуры входных и выходных данных.  Разработка алгоритма решения задачи. Определение формы представления входных и выходных данных.  Определение семантики и синтаксиса языка.  Разработка структуры программы. Окончательное определение конфигурации технических средств. |
| Утверждение технического проекта | Разработка плана мероприятий по разработке и внедрению программ. Разработка пояснительной записки. |

|  |  |  |
| --- | --- | --- |
| Стадии разработки | Этапы работ | Содержание работ |
|  |  | Согласование и утверждение технического проекта. |
| 4. Рабочий проект | Разработка программы | Программирование и отладка программы. |
| Разработка программной документации | Разработка программных документов в соответствии с требованиями ГОСТ 19.101-77. |
| Испытания программы | Разработка, согласование и утверждение порядка и методики испытаний.  Проведение предварительных государственных, межведомственных, приёмо-сдаточных и других видов испытаний.  Корректировка программы и программной документации по результатам испытаний. |

Продолжение таблицы 2

1. **Порядок контроля и приемки**

В качестве испытаний проверяется корректность работы программы на сервере. Испытания заключаются в функциональном тестировании системы и проверку установленных зависимостей.

Система считается готовой к эксплуатации после успешного проведения испытаний.

1. **Приложения**

В приложении А приведено полное техническое задание.