МИНИСТЕРСТВО ОБРАЗОВАНИЯ КИРОВСКОЙ ОБЛАСТИ
Кировское областное государственное профессиональное
образовательное бюджетное учреждение
«Слободской колледж педагогики и социальных отношений»

Дипломный проект допущен к защите
Заместитель директора по воспитательной и методической работе
_____ к.п.н., Глазырина Т. Г.
«_____»_____ 2024 г.

ДИПЛОМНЫЙ ПРОЕКТ

РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ БОЛЬШИХ ОБЪЕМОВ ТЕКСТОВЫХ ДАННЫХ

Выполнила
Поглазова Валерия Владимировна
студент специальности 09.02.07
Информационные системы
и программирование
группа 21П-1
Форма обучения: очная

_____
(подпись)

Руководитель:
Калинин Арсений Олегович

_____
(подпись)

Дипломный проект защищен
«_____»_____2024 г.
Оценка _____
Секретарь ГЭК_____

Слободской
2024

Нормоконтроль:

Дата: _____

_____ _____

Подпись Расшифровка подписи

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В наше время наблюдается стремительный рост объемов цифровых данных, значительная часть которых представлена в текстовом формате. Эти данные хранятся в различных форматах и источниках: от документов и веб-страниц до социальных сетей и баз данных. Однако, эта информация часто является неструктурированной или полуструктурированной, что затрудняет её анализ и использование для принятия решений. Извлечение полезной информации из таких массивов данных вручную является трудоемким и неэффективным процессом.

Актуальность данной темы заключается в том, что эффективное извлечение информации из больших объемов текстовых данных является критически важной задачей для многих областей, включая науку, бизнес, государственное управление и другие. Автоматизация этого процесса позволяет значительно сократить время и затраты на анализ данных, повысить точность и объективность результатов, а также открыть новые возможности для принятия информированных решений.

В процессе проведения исследования выявляется, что существующие решения для обработки больших текстовых данных часто обладают недостаточной гибкостью, не поддерживают широкий спектр форматов данных или имеют ограниченную функциональность. Многие инструменты требуют значительных вычислительных ресурсов или обладают высокой стоимостью.

Разработка специализированных программ для извлечения информации из больших объемов текстовых данных становится все более актуальной задачей в связи с постоянным увеличением количества цифровых данных, ростом сложности задач анализа данных и необходимостью повышения эффективности работы с информацией. Это особенно важно в условиях, когда требуется обработка данных из различных источников и в разных форматах.

Таким образом, разработка программы "Извлечение информации из

больших объемов текстовых данных" решает конкретные проблемы неэффективности и трудоемкости ручного анализа текстовых данных, но и подчеркивает актуальность автоматизации процессов обработки больших данных. Программа позволит эффективно извлекать, структурировать и анализировать информацию из различных источников, что существенно повысит производительность и качество работы с данными. Поэтому данную программу целесообразно написать на языке программирования C#, учитывая его широкие возможности для работы с данными, наличие развитых библиотек для обработки текста и высокую производительность.

Язык программирования C# предоставляет разработчику поистине великолепный набор простых в использовании инструментов, позволяющих быстро разрабатывать сложные проекты, создавая приятный и удобный пользовательский интерфейс.

Объект исследования: процесс извлечения структурированной информации из больших объемов неструктурированных текстовых данных.

Предмет исследования: разработка программного обеспечения для извлечения информации из больших объемов текстовых данных

Цель дипломного проекта – заключается в разработка и реализации программного обеспечения, способного эффективно извлекать заданную информацию из больших объемов текстовых данных, преобразовывать ее в удобный для анализа и дальнейшей обработки структурированный формат, обеспечивая высокую скорость и точность извлечения.

Задачи исследования:

- Описать предметную область.
- Разработать технического задание на создание программного продукта.
- Описать архитектуру программы.
- Описать алгоритмы и функционирование программы.
- Провести тестирование и опытную эксплуатацию.

- Разработать руководство оператора

Практическая значимость: заключается в создании инструмента, позволяющего автоматизировать и ускорить процесс извлечения и структурирования информации из больших объемов текстовых данных.

Методы исследования - системный анализ и функциональное моделирование.

ГЛАВА 1. АНАЛИТИЧЕСКАЯ ЧАСТЬ

1.1. Анализ предметной области

Извлечение информации из больших объемов текстовых данных (часто сокращается до IEBD - Information Extraction from Big Data)— это процесс автоматического извлечения структурированной информации из неструктурированных или полуструктурированных текстовых данных.

Извлечение информации из больших объемов текстовых данных разрабатывалась для автоматизации процесса анализа и извлечения ключевой информации из неструктурированных и полуструктурированных текстовых данных, хранящихся в различных форматах и источниках (документы, веб-страницы, социальные сети и т.д.).

Этот процесс был осуществлен с целью повышения эффективности и производительности обработки больших объемов текстовых данных, снижения трудозатрат на ручной анализ, повышения точности и объективности результатов анализа, а также обеспечения возможности принятия более обоснованных решений на основе полученной информации.

Благодаря извлечению информации из больших объемов текстовых данных можно обнаружить скрытые закономерности и тренды, улучшить принятие решений в бизнесе, науке и государственном управлении, оптимизировать процессы обработки информации, получить конкурентное преимущество за счет быстрого доступа к важной информации, автоматизировать мониторинг СМИ и социальных сетей.

Извлечение информации из больших объемов текстовых данных представляет собой комплексный процесс, включающий в себя этапы загрузки данных, предобработки, извлечения ключевых сущностей, анализ настроений, структурирование данных и визуализации данных.

Цель извлечения состоит в том, чтобы преобразовать неструктурированные или полуструктурированные текстовые данные в структурированный формат, удобный для дальнейшего анализа и использования, извлечь из них конкретную информацию и выявить закономерности, необходимые для решения конкретных задач пользователя.

Также в процессе анализа было проведено исследование предметной области, в результате которого была создана диаграмма вариантов использования, которая отражает актера (пользователя) и выполняемые им функции (Рисунок 1.1).

На ней видно, что в системе выделяется несколько центральных пользователей:

- Пользователь – может загружать текстовые данные из различных источников и форматов (файлы, URL); сохранять результаты анализа в различных форматах; экспортировать результаты в различные форматы (CSV, JSON, XML) для дальнейшей обработки в других системах.

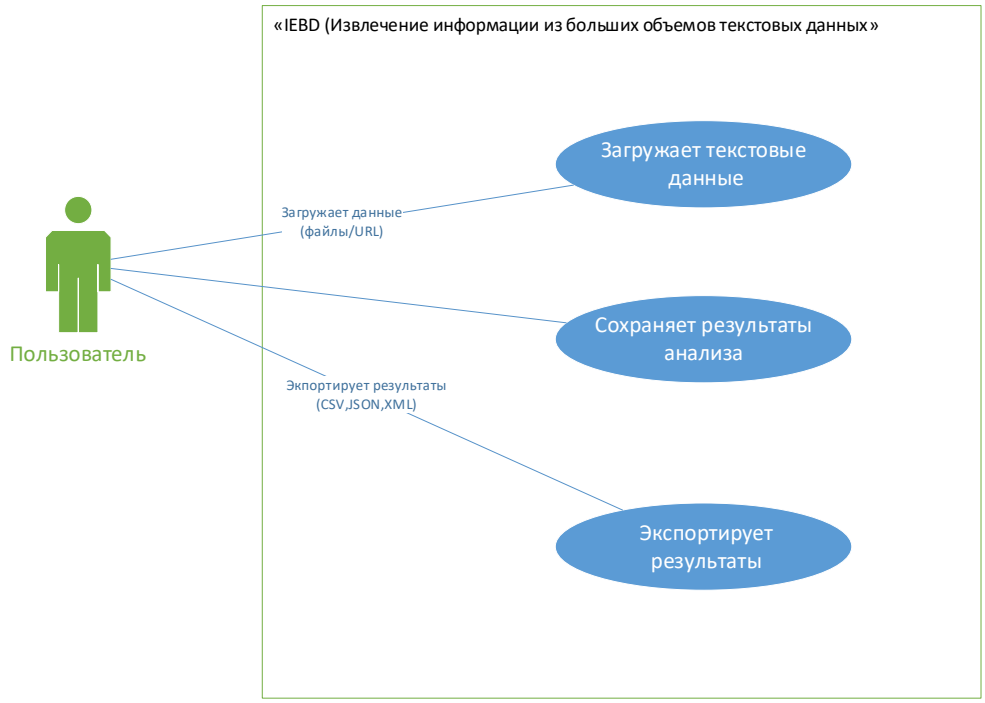

Рисунок 1.1 - Диаграмма вариантов использования

Анализ программы аналога.

Postman — это платформа для разработки и тестирования API. Она представляет собой интегрированную среду разработки (IDE), которая позволяет разработчикам создавать, отправлять и тестировать HTTP-запросы, управлять коллекциями запросов, автоматизировать тестирование, и получать данные о производительности API.

Плюсы Postman:

1. Удобный интерфейс: интуитивно понятный и простой в использовании интерфейс, даже для начинающих разработчиков.
2. Автоматизация запросов: Postman позволяет автоматизировать отправку запросов к веб-сайтам и API, что может сэкономить время при извлечении больших объемов данных.
3. Многофункциональность: Postman поддерживает множество типов запросов (GET, POST, PUT, DELETE и др.), а также различные форматы данных (JSON, XML, и другие). Поддерживает автодополнение, что ускоряет создание запросов.
4. Экспорт данных: Postman позволяет экспортировать полученные данные в различные форматы, такие как CSV и JSON, что упрощает дальнейший анализ.
5. Бесплатная версия с достаточным функционалом: доступна бесплатная версия, которая вполне подходит для многих задач.

Минусы Postman:

1. Ограничения бесплатной версии: бесплатная версия имеет некоторые ограничения по количеству запросов, коллекций и пользователей.
2. Цена платной версии: платная версия может быть довольно дорогой для отдельных разработчиков или небольших команд.
3. Зависимость от интернета: для работы Postman требуется подключение к интернету.
4. Сложность для очень сложных сценариев: для очень сложных и многоуровневых API-сценариев или работы с необычными форматами данных,

Postman может показаться несколько ограниченным, особенно в сравнении с инструментами программирования.

5. Ограничения в обработке больших объемов данных: не самый подходящий инструмент для очень больших объемов данных из-за его фокуса на взаимодействие с API и работе с небольшими наборами данных.

6. Слишком много возможностей: для новичка может показаться перегруженным из-за большого набора инструментов и функций, которые не всегда нужны.

1.2. Техническое задание

Техническое задание разрабатывалось на основании заявки (Приложение 1) и требований заказчика, программы «Извлечение информации из больших объемов текстовых данных», анализа предметной области и с учетом требований ГОСТ 19.201-78 [5].

Наименование программы – «IEBD (Извлечение информации из больших объемов текстовых данных». Программа предназначена для извлечения информации из больших объемов текстовых данных.

Разработка программы ведется на основании учебного плана и перечня тем утвержденных на заседании предметно цикловой комиссии информатики и программирования.

Функциональным назначением программы является извлечение информации из больших объемов текстовых данных.

Программа должна обеспечивать возможность выполнения перечисленных ниже функций:

- Загрузка данных;
- Предобработка данных (Очистка данных: Удаление лишних символов, специальных символов и т.д.);
- Извлечение информации;

- Формирование структурированных данных;
- Сохранение результатов.

Надежное (устойчивое) функционирование программы должно быть обеспечено выполнение заказчиком совокупности организационно-технических мероприятий, перечень которых приведен ниже:

- организация бесперебойного питания технических средств;
- использование лицензионного программного обеспечения;
- отсутствие вредоносного программного обеспечения, наличие антивирусной программы;
- соблюдение правил и требований по эксплуатации технических средств.

Время восстановления после отказа, вызванного сбоем электропитания технических средств (иными внешними факторами), не фатальным сбоем (не крахом) операционной системы, не должно превышать 5 минут при условии соблюдения условий эксплуатации технических и программных средств.

Время восстановления после отказа, вызванного неисправностью технических средств, фатальным сбоем (крахом) операционной системы, не должно превышать времени, требуемого на устранение неисправностей технических средств и переустановки программных средств.

Отказы программы возможны вследствие некорректных действий оператора (пользователя) при взаимодействии с операционной системой. Во избежание возникновения отказов программы по указанной выше причине следует обеспечить работу пользователя без предоставления ему административных привилегий.

Климатические условия эксплуатации, при которых должны обеспечиваться заданные характеристики, должны удовлетворять требованиям, предъявляемым к техническим средствам в части условий их эксплуатации.

В состав технических средств должен входить IBM-совместимый персональный компьютер (ПЭВМ), включающий себя:

- процессор с тактовой частотой, 1 ГГц, не менее;
- оперативную память объемом 512 Мб, не менее;
- жесткий диск со свободным местом 500 Мб, не менее;
- монитор, с разрешением экрана 1024*768, не менее;
- компьютерная мышь;
- клавиатура;
- принтер.

Исходные коды программы должны быть реализованы на языке C#. В качестве интегрированной среды разработки программы должна быть использована среда программирования Microsoft Visual Studio 2022.

Системные программные средства, используемые программой, должны быть представлены лицензионной локализованной версией операционной системы Windows 7/8/10/11.

Программное обеспечение поставляется в виде изделия на USB-флэш накопителе.

Требования к транспортировке и хранению должны соответствовать условиям эксплуатации носителей, на которых находится программный продукт.

Программа должна обеспечивать взаимодействие с пользователем посредством графического пользовательского интерфейса.

Предварительный состав программной документации включает в себя следующие документы:

- техническое задание;
- руководство оператора.

Разработка должна быть проведена в следующие стадии и этапы:

1. Анализ требований:

На стадии анализ требований формулируются цели и задачи проекта. Создается основа для дальнейшего проектирования

2. Проектирование:

На стадии проектирование должны быть выполнены перечисленные ниже этапы работ:

- разработка программной документации;

На этапе разработка программной документации должна быть выполнена разработка технического задания.

При разработке технического задания должны быть выполнены перечисленные работы: постановка задачи, определение и уточнение требований к техническим средствам, определение требований к программе, определение стадий, этапов и сроков разработки программы и документации на нее, выбор языков программирования.

- разработка алгоритма программы;

На этапе разработки алгоритма программы должен быть разработан алгоритм работы программы.

- кодирование;

На стадии кодирования происходит реализация алгоритмов в среде программирования.

- тестирование и отладка.

На стадии тестирование и отладка происходит проверка алгоритмов, реализованных в программе на работоспособность в различных ситуациях. Исправление выявленных ошибок, повторное тестирование.

Приемо-сдаточные испытания должны проводиться при использовании технических средств. Приемка программы заключается в проверке работоспособности программы путем ввода реальных или демонстрационных данных.

Во время приемки работы разработчик предоставляет программу и документацию, которая к ней прилагается. Проводятся испытания программы, при успешных испытаниях программа вводится в эксплуатацию. При ошибках, недопустимых для успешной работы программного продукта – отправляется на доработку.

Было описано техническое задание, содержащее в себе информацию о программном продукте, его функциях, эксплуатации и требования, которые должны учитываться при создании программы и документации к ней.

ВЫВОДЫ ПО ГЛАВЕ 1

В этой главе была описана аналитическая часть, содержащая в себе описание извлечения информации из больших объемов текстовых данных, который был изучен перед созданием программного продукта, а также техническое задание, содержащее в себе информацию о программном продукте, его функциях, эксплуатации и требования, которые должны учитываться при создании программы и документации к ней.

Определена значимость и требования к будущему программному обеспечению. В техническом задании были определены основные требования к программному продукту и функциональные характеристики, а также состав программной документации.

ГЛАВА 2. КОНСТРУКТОРСКАЯ ЧАСТЬ

2.1. Архитектура программы

Разработанный программный продукт поставляется в виде установочного файла «mysetup.exe».

На рисунке 2.2 показана схема программы. Данная схема отображает алгоритм работы программы.

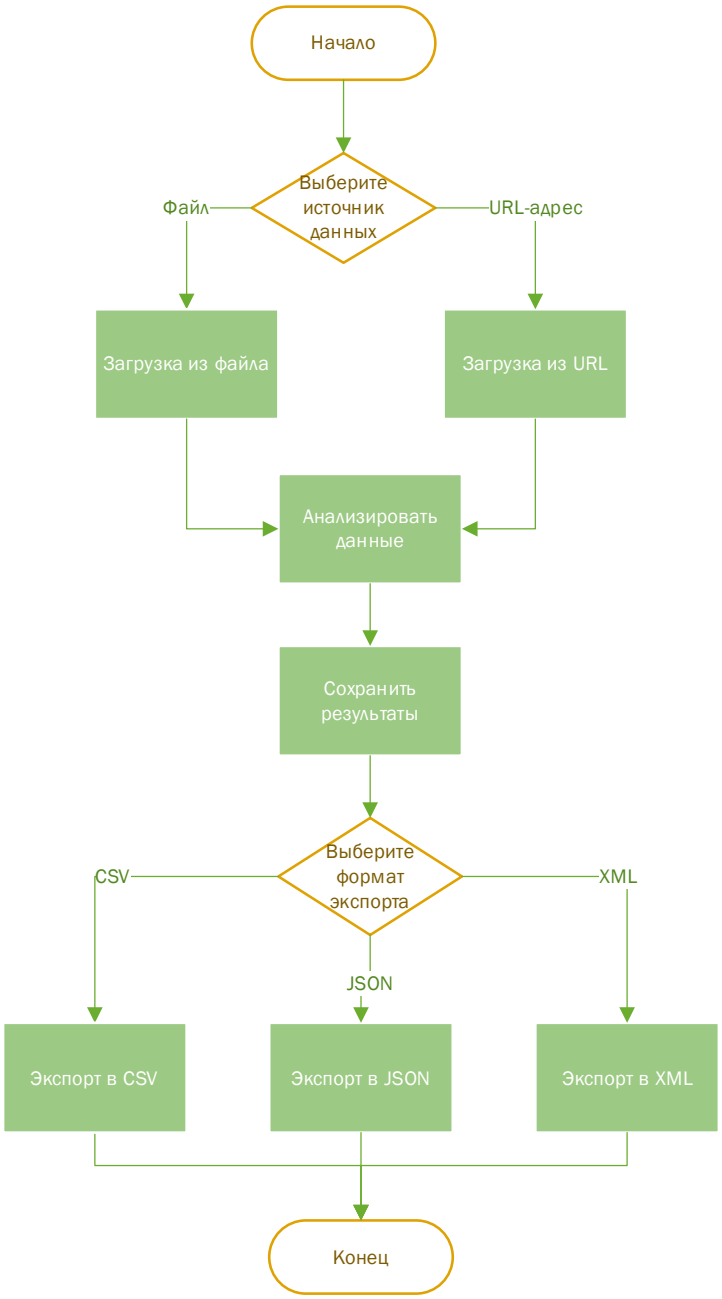

Рисунок 2.2 – Алгоритм работы программы

Весь код программы описан в разделе приложений (Приложение 2).

Приложение «IEBD (Извлечение информации из больших объемов текстовых данных», написанное на языке программирования C# в интегрированной среде разработки Microsoft Visual Studio 2022, работает под управлением операционной системы Windows 11.

В программном продукте реализовано одно главное окно. Окно предоставляет основной функционал программы, и вся работа происходит именно в нем. В основном окне реализованы следующие функции:

- Текстовое поле «Введите URL-адрес или вставьте текст» предназначено для ввода пользователем URL-адреса веб-страницы или файла, который требуется обработать;
- Кнопка «Открыть файл» предназначена для предоставления пользователю возможности выбора файла с локального диска для обработки.
- Кнопка «Отправить» предназначена для отправки введенного URL-адреса или файла в текстовом поле для извлечения данных;
- Список предназначен для отображения результатов обработки введенных данных;
- Выпадающий список предназначен для выбора формата файла, в который пользователь хочет сохранить результаты обработки. Доступные форматы: XML, JSON и CSV;
- Кнопка «Сохранить файл» предназначена для сохранения результатов обработки (отображаемых в списке) в файл выбранного формата (выбранного в выпадающем списке);

2.2. Описание алгоритмов и функционирования программы

Функционирование программы

Алгоритмы программы

ВЫВОДЫ ПО ГЛАВЕ 2

В результате над конструкторской частью была разработана архитектура созданного программного продукта, описаны алгоритмы и функционирования программы.

Для более наглядного понимания архитектуры программы и алгоритма её работы была построена схема работы программного обеспечения.

Вся информация представлена для изучения структуры программы и ее работы.

ГЛАВА 3. ЭКСПЕРИМЕНТАЛЬНО-ПРИКЛАДНАЯ ЧАСТЬ

3.1. Тестирование и опытная эксплуатация программы

3.2. Руководство оператора

Функциональным назначением программы является извлечение информации из больших объемов текстовых данных.

Программа должна обеспечивать возможность выполнения перечисленных ниже функций:

- Загрузка данных;
- Предобработка данных (Очистка данных: Удаление лишних символов, специальных символов и т.д.);
- Извлечение информации;
- Формирование структурированных данных
- Сохранение результатов;

Условия выполнения программы

Климатические условия эксплуатации, при которых должны обеспечиваться заданные характеристики, должны удовлетворять требованиям, предъявляемым к техническим средства в части условий их эксплуатации.

Минимальный состав технических средств

В состав технических средств должен входить IBM-совместимый персональный компьютер (ПЭВМ), включающий в себя:

- процессор с тактовой частотой, 1 ГГц, не менее;
- оперативную память объемом, 512 Мб, не менее;
- жесткий диск со свободным местом 500 Мб, не менее;
- монитор, с разрешением экрана 1024*768, не менее;
- компьютерная мышь;
- клавиатура;
- принтер;

Минимальный состав программных средств.

Системные программные средства, используемые программой, должны быть представлены лицензионной локализованной версией операционной системы Windows 10 и выше.

Требования к персоналу (пользователю).

Минимальное количество персонала требуемого для работы программы, должно составлять не менее 1 штатной единицы – пользователь программы, оператор.

Пользователь программы должен обладать практическими навыками работы с графическим пользовательским интерфейсом операционной системы семейства Windows.

Загрузка и запуск программы

Для установки программы необходимо открыть файл «mysetup.exe» от имени администратора с USB-флэш накопителя (Приложение 3).

ВЫВОДЫ ПО ГЛАВЕ 3

ЗАКЛЮЧЕНИЕ

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ 19.103-77. Единая система программной документации. Обозначение программ и программных документов, введ. 01.07.1978. – М.: Изд-во стандартов, 1978. – 2 с.
2. ГОСТ 19.104-78. Единая система программной документации. Основные надписи, введ. 01.01.1978. – М.: Изд-во стандартов, 1980. – 6 с.
3. ГОСТ 19.106-78. Единая система программной документации. Требования к программным документам, выполненным печатным способом, введ. 01.01.1980. – М.: Изд-во стандартов, 1989. – 6 с.
4. ГОСТ 19.201-78. Единая система программной документации. Техническое задание. Требования к содержанию и оформлению, введ. 01.01.1980. – М.: Изд-во стандартов, 1988. – 3 с.
5. ГОСТ 19.401-78. Единая система программной документации. Текст программы. Требования к содержанию и оформлению, введ. 01.01.1980. – М.: Изд-во стандартов, 1988. – 2 с.
6. ГОСТ 19.505-79. Единая система программной документации. Руководство оператора. Требования к содержанию и оформлению, введ. 01.01.1980. – М.: Изд-во стандартов, 1987. – 6 с.

ПРИЛОЖЕНИЯ

Приложение 2

Код приложения

MainWindow.xaml

```
<Window x:Class="Inv.MainWindow"

xmlns="http://schemas.microsoft.com/winfx/2006/xaml/presentation"

xmlns:x="http://schemas.microsoft.com/winfx/2006/xaml"

xmlns:d="http://schemas.microsoft.com/expression/blend/2008"

xmlns:mc="http://schemas.openxmlformats.org/markup-compatibility/2006"

xmlns:materialDesign="http://materialdesigninxaml.net/winfx/xaml/themes"

    xmlns:local="clr-namespace:Inv"

    mc:Ignorable="d"

    Height="550" Width="1100"

    MinHeight="400"

    MinWidth="800"

    x:Name="WindowChrome"

    Title="Учёт спортивного инвентаря | Авторизация"

    Icon="icon.ico"

    Style="{StaticResource VS2012WindowStyle}" Loaded="Window_Loaded" Closed="WindowChrome_Closed">

  <Window.Resources>

    <ResourceDictionary>

      <ResourceDictionary.MergedDictionaries>

        <materialDesign:BundledTheme BaseTheme="Light" PrimaryColor="Blue" SecondaryColor="LightGreen" ColorAdjustment="{materialDesign:ColorAdjustment}"/>

      </ResourceDictionary.MergedDictionaries>

    </ResourceDictionary>

  </Window.Resources>

  <Grid>

    <!-- MENU -->

    <Grid x:Name="AllGrid">

      <Grid.ColumnDefinitions>

        <ColumnDefinition MinWidth="180" MaxWidth="180"/>

        <ColumnDefinition Width="3*"/>

      </Grid.ColumnDefinitions>

      <Grid Grid.Column="0">

        <ScrollViewer Style="{StaticResource LeftScrollViewer}">

          <StackPanel>

            <Grid Height="10">

              <Grid.ColumnDefinitions>

                <ColumnDefinition Width="11*"/>

                <ColumnDefinition Width="0*"/>

                <ColumnDefinition Width="4*"/>

              </Grid.ColumnDefinitions>

            </Grid>

            <Button x:Name="MenuBtn2" Style="{StaticResource MenuBtn}" Click="MenuBtn2Click" Content="Учёт"/>

            <Button x:Name="MenuBtn1" Click="MenuBtn1Click" Style="{StaticResource MenuBtn}" Content="Инвентарь"></Button>

            <Button x:Name="MenuBtn7" Click="MenuBtn7Click" Style="{StaticResource MenuBtn}" Content="Списанное"></Button>

            <Button x:Name="MenuBtn3" Click="MenuBtn3Click" Style="{StaticResource MenuBtn}" Content="Экспорт в Word"/>

            <Button x:Name="MenuBtn4" Click="MenuBtn4Click" Style="{StaticResource MenuBtn}" Content="Экспорт в Excel"/>
```

Приложение 3

USB-флэш накопитель с материалами проекта

На USB-флэш накопителе располагается:

- Проект приложения (каталог Inv)
- Установщик приложения (файл mysetup.exe)
- Файл дипломного проекта в формате MS Word (Дипломный проект.docx)