ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет информатики, математики и компьютерных наук

Программа подготовки бакалавров по направлению

38.03.05 Бизнес-информатика

Пилипенко Глеб Игоревич

Курсовая работа

по теме “Продуктизация и доработка навыка голосового помощника Алиса для работы с расписанием и новостями ВШЭ”

Выполнил:

Пилипенко Глеб Игоревич 22БИ-3

Принял: Дыдычкин Александр Алексеевич

**Нижний Новгород, 2024**

**Оглавление**

[Пояснительная записка 3](#_heading=h.gjdgxs)

[Техническое задание 7](#_heading=h.1fob9te)

[Анализ проблематики 10](#_heading=h.3znysh7)

[Анализ существующих решений, технологий и аналогов 16](#_heading=h.2et92p0)

[Полученные результаты, методика испытаний и итоги анализа 23](#_heading=h.tyjcwt)

[Список использованных источников 31](#_heading=h.3dy6vkm)

**Пояснительная записка**

**Введение**

Наименование: Написание модуля по внесению информации о студентах, преподавателях Высшей школы экономики, их расписании и работа с этими данными в Yandex Data Base, проектирования подпрограммы по взаимодействию с пользователем во время сессии навыка..

Разработка ведется на основании данный документов:

1. Официальная документация языка программирования Python.
2. Официальная документация по YDB.
3. Официальная документация YDB Python SDK.

Современные голосовые помощники становятся неотъемлемой частью повседневной жизни благодаря своей простоте и функциональности. Среди множества виртуальных ассистентов, представленных на российском рынке, особой популярностью пользуется «Алиса» от компании Яндекс. Данный голосовой помощник согласно [данным](https://sprut.ai/news/chto-proishodit-na-rynke-golosovyh-assistentov-v-rossii) доля, занимаемая данным продуктом на рынке выросла с 44% в октябре 2020 года до 83% в августе 2024.

Российский рынок умных колонок в 2024 году показал значительный рост. По данным группы «М.Видео-Эльдорадо», опубликованным 17 января 2025 года в TAdviser, объем продаж таких устройств увеличился на 20% в штучном выражении и на 26% в денежном по сравнению с предыдущим годом.

Цель данной работы – оптимизировать взаимодействие пользователя с голосовым помощником при получении информации о расписании, сделав этот процесс максимально быстрым и комфортным. Также необходимо пересмотреть выполненное ранее решение, оптимизировать его, сделать более универсальным и доступным для пользователей не только филиала НИУ ВШЭ в Нижнем Новгороде, но и в остальных кампусах университета.

Проблематика данной курсовой работы заключается в использовании баз данных компании Яндекс для удовлетворения потребности в универсальности и добавления возможности пользователю вносить данные расписания самостоятельно. Разрабатываемая подпрограмма в рамках командного проекта выполняется функцию по внесению, сбору, изменению и хранению информации о пользователях. Также выполнена работа по составлению прототипа модуля взаимодействия с пользователем.

**Назначение и область применения**

Цель реализуемого программного модуля заключается во внесении информации о пользователе, его расписании, поиску внесенного расписания в базе данных и приведении его в читаемый для пользователя вид.

Область применения включает в себя, в основном, использование студентами и преподавателями НИУ ВШЭ в независимости от кампуса при использовании голосового ассистента Алиса или применение в рамках других работ, где существует необходимость внесения расписания и информации в базу данных.

**Технические характеристики**

Главной задачей являлось анализ и изменение кода, который был выполнен в прошлой работе, а именно: изменение логики модуля взаимодействия с пользователем, добавление преподавательского режима, возможности вносить и получать пары майноров, английского и факультативов, расширение области применимости, которая бы включала все кампусы НИУ ВШЭ, написание подпрограммы, которая будет по полученным данным из модуля взаимодействия с пользователем осуществлять внесение информации о нем в базу данных, а также о предметах, которые необходимо добавить и осуществлять поиск по базе данных выбранных предметов и предоставлять массив, состоящих их объектов базы данных, которые содержат информацию о дате и времени занятий, наименовании предмета, номере аудитории, фамилии преподавателя и тип пар, т.е. лекционное, практическое занятие или научно-исследовательский семинар.

К основным ограничением данного способа хранения и заполнения расписания можно отнести: сложность в составлении и унификации списка предметов, т.е. заранее необходимо знать список дисциплин всех образовательных программ НИУ ВШЭ, для их обработки, необходимость постоянно добавлять информацию о дисциплинах вручную, что может быть не удобно пользователю.

Методы, которые применялись для разработки навыка включали использование официальных библиотек YDB и самого сервиса для создания заранее спроектированной базы данных, заполнения необходимой информации о пользователе: фамилия, имя, отчество и номер группы, если это студент, информации о группе, расписании. Используемый способ подразумевает в независимости от кампуса, вида предмета хранить расписание, что делает его более гибким, чем решение с парсингом гугл таблиц, а также в виду того, что все предметы хранятся в заранее установленным виде, универсальность данного решения также выше. Подпрограмма состоит из пяти модулей:

1. Модуль models, который содержит классы данных: базовый класс Lesson для занятий, содержащий атрибуты name, type\_l, building, auditorium, id\_lecturer, time, is\_weekly, is\_upper, lesson\_date. Два класса наследника PersonalLesson**,** GroupLesson, которые наследуют Lesson, добавляя соответственно атрибуты  id\_student или id\_group. Также присутствуют вспомогательные классы для таблиц: PersonalLessonStudent, GroupLessonGroup. Помимо этого в данном модуле содержатся классы Student, Lecturer для пользователей, содержащие данные о ФИО и номер группы, если это студент, объект группа, который хранит атрибуты name, edu\_year, edu\_program, faculty, edu\_format, edu\_level и вспомогательный класс LecturerGroup.
2. Модуль obj\_queries, содержащий функции для каждого из объектов: Group, Student, Lecturer по получению записей в базе данных объекта, идентификатора и проверке регистрации.
3. Модуль Registration\_ydb, хранящий в себе общую функцию регистрации registration\_user и методы, которые используются ей в зависимости от параметров для регистрации студента, лектора или группы.
4. Модуль schedule\_queries, осуществляющий внесение расписания в таблицы: insert\_lesson, insert\_lesson\_data insert\_help\_tables\_data, его поиск и изменение данных пользователя в БД. В силу особенностей хранения данных и выполнения запросов функции для поиска предмета по названию, лектору, дате разделены для студентов и преподавателей и также присутствуют отдельные методы для нахождения предмета по дню недели: find\_by\_week\_day\_lesson\_student, find\_by\_week\_day\_lesson\_lecturer.
5. Модуль utils содержит функцию make\_readable для вывода данных пользователю в заданном формате.

Описанная выше часть подпрограммы будет осуществлять взаимодействие с модулем по обработке пользовательских запросов во время работы навыка голосового ассистента, поэтому данный алгоритм преобразовывает данные, хранящиеся в YDB в массив атрибутов row класса ydb.ResultSet. Данный тип данных можно удобно использовать и вычленять всю необходимую информацию.

Входные данные содержат объект ydb.QuerySessionPool, который необходим для каждой функции по взаимодействию с базой данных и варьируются в зависимости от задачи. Так методы поиска требуют указания параметров, по которым будет осуществляться поиск, связан ли предмет с данной образовательной программой, идентификаторов студента, группы. В функции добавления предмета требуются данные пользователя в виде списка. Также некоторые другие функции требуют передачи объектов Lesson, Student или Lecturer в зависимости от задачи.

Написанная подпрограмма выполнена на языке программирования Python. Можно выделить следующие основные моменты, которые повлияли на данное решение:

* Использование остальными участниками данного языка программирования, поэтому для интеграции рассматриваемого решения необходимо применение единственного решения.
* Высокая популярность рассматриваемого языка программирования.
* YDB обладает обширной документацией и множеством готовых библиотек, что значительно упрощает работу с этим инструментом. Кроме того, выбранный язык программирования обеспечивает высокую скорость разработки благодаря интуитивному синтаксису, активной поддержке комьюнити и постоянному развитию экосистемы.
* Важно и то, что этот язык широко применяется для обработки и анализа данных, включая взаимодействие с YDB, а также с другими источниками информации, что делает его универсальным решением для подобных задач.

Разработанный программный модуль размещается в облачной инфраструктуре Yandex Cloud. Эта платформа предлагает готовые решения для хранения и выполнения кода, специально адаптированные для работы с Яндекс Диалогами. Особенно стоит отметить возможность бесплатного тестирования голосовых навыков и их размещение через сервис Yandex Cloud Functions. Помимо этого, применяемые технологии включают в себя использование сервиса YDB – распределенная реляционная система управления базами данных с открытым исходным кодом на основе языка SQL. Данное информационное решение является ключевым в разрабатываемой программе, потому что оно поддерживают удобную и понятную интеграцию с сервисом Yandex Cloud Functions, в котором размещен сам навык и выполняет функцию по хранению и обработке информации о пользователях и их расписании. Также одним из преимуществ данного продукта является возможность бесплатного использования в рамках ограниченных лимитов.

Применение указанных инструментов существенно оптимизирует работу, сокращая время на тестирование, разработку и внедрение модуля по хранению, извлечению и работе с данными студентов и преподавателей и их расписания.

**Техническое задание**

Модуль по работе с данными пользователей - программа, которая позволяет осуществлять заполнение, хранение, поиск и дальнейшую обработку информации, полученной от студентов и преподавателей в процессе взаимодействия с голосовым ассистентом.

Цель в рамках описанного продукта является создание программного модуля, который будет осуществлять вышеописанные действия с информацией о пользователях всех кампусов НИУ ВШЭ, обеспечивая быстрый и точный доступ к информации через голосовые и текстовые запросы.

Разрабатываемая подпрограмма представляет собой ключевой компонент системы, реализованной в виде навыка для голосового помощника Алиса. Данное программное решение призвано усовершенствовать взаимодействие пользователей с системой, предоставив новый пользовательских опыт.

Основное преимущество реализации заключается в создании интуитивно понятного интерфейса голосового управления, который обеспечивает быстрый и комфортный доступ к запрашиваемой информации. Такой подход существенно сокращает временные затраты пользователей по сравнению с традиционными методами поиска данных, одновременно повышая точность и релевантность получаемых результатов.

**Функциональные требования:**

1. Разрабатываемый программный модуль должен взаимодействовать с продуктом YDB для работы с информацией о студентах и преподавателях различных программ и форматов обучения Высшей школы экономики, их расписании.
2. Программный модуль должен вносить информацию о пользователях в БД, а именно фамилию, имя, отчество и номер группы, если пользователь студент. Также должна фиксироваться информация о группе: название, курс, образовательная программа, факультет, формат обучения, уровень образования. Помимо того, фиксируются данные о предмете, которые включают название, тип предмета (семинар или лекция), корпус, аудитория, преподаватель, время, дата и группа, если предмет проводится для конкретной образовательной программы, или студент, если данный предмет является индивидуальным для студента, например, майнор, факультатив.
3. Программный модуль должен извлекать необходимые данные о расписании пользователей по атрибутам: название, лектор, дата, которая включает, как конкретное число и месяц, так и определенный день недели или пары на сегодня и завтра.
4. Собранные сведения требуется записать в стандартизированном формате, обеспечивающем их дальнейшее использование.
5. При необходимости пользователь должен иметь возможность изменить хранящуюся в базе данных информацию.
6. Разрабатываемый программный модуль должен осуществлять регистрацию преподавателей или групп, если они указываются в атрибутах предмета, но не содержатся в базе данных.

**Требования к надежности:**

1. Программный компонент обязан предусматривать корректную реакцию на ситуации разрыва соединения с СУБД, включая механизмы повторного подключения и обработки исключительных состояний
2. Компонент работы с расписанием должен демонстрировать эффективное использование ресурсов при получении данных минимальное время отклика и достаточную производительность для соответствия требованиям по времени ответа голосового ассистента.

**Требования к информационной и программной совместимости:**

1. Язык программирования: Python (минимальная поддерживаемая версия — 3.12). Независимость от платформы: функционирование должно быть обеспечено вне зависимости от операционной системы или аппаратной конфигурации устройства, на котором работает голосовой помощник.
2. Модуль должен обеспечивать корректную работу с библиотекой по взаимодействию с СУБД YDB.
3. Итоговое программное решение должно быть совместимо с другими модулями, предусмотренными в проекте, а именно получать на вход ссылку и возвращать на выходе данные для модуля поиска необходимых элементов расписания.
4. Программа должна быть реализована с использованием принципов ООП.

**Требования к маркировке и упаковке**

Финальная версия приложения подлежит публикации в официальном каталоге навыков голосового помощника Яндекса. Что касается модуля обработки и хранению данных в YDB, специальные требования к его маркировке и способу упаковки не устанавливаются.

**Требования к транспортированию и хранению.**

Подпрограмма по работе с информацией и пользователях и их расписании может располагаться на виртуальной машине, поддерживающей использование и развертывание Webhook сервера и взаимодействие с YDB, бессерверной платформе Yandex Cloud Funtions или в другом месте, которое соответствуют требованиям выше как часть общего навыка голосового ассистента до момента внесения изменений разработчиками или работниками платформы. К транспортировке никакие требования не предъявляются.

**Тестирование**

1. Нужно проверить разработанную программу при различных входных значениях, включающих ситуации, когда пользователь является не только студентом, но и преподавателем, отсутствует регистрация группы или преподавателя при вводе информации о предмете, вводятся различные данные предмета.
2. Необходимо сделать проверку выполнения CRUD-операций, т.е. создание новых записей, их получение и передача в нужном формате, обновление строк таблицы и их удаление.
3. Протестировать корректную передачу данных подмодулю по обработке запросов пользователя и скорость выполнения запросов, чтобы она не превышала заданного времени для ответа навыка.

**Стадии и этапы разработки**

1. Формирование технического задания:

-Анализ и систематизация функциональных характеристик подпрограммы

-Определение нефункциональных параметров реализации

1. Исследование исходных данных:

-Анализ необходимой структуры данных для хранения

-Проектирование базы данных, включающее создание таблиц и связей между ними

1. Создание архитектуры:

-Разработка структурной схемы модуля

-Декомпозиция на ключевые компоненты и интерфейсы взаимодействия

1. Реализация базового функционала:

-Программная интеграция с сервисом YDB.

-Создание алгоритмов внесения, извлечения и трансформации информации.

1. Верификация работоспособности:

-Тестирование корректности обработки данных

-Проверка устойчивости к вариациям входных форматов

1. Отладка и оптимизация:

-Устранение выявленных ошибок

-Повышение надежности работы модуля

1. Обеспечение совместимости:

-Адаптация интерфейсов для интеграции с другими компонентами навыка

-Тестирование кроссплатформенной работы

1. Финальная интеграция:

-Встраивание модуля в общую систему

-Комплексное тестирование в составе проекта

-Корректировка взаимодействующих компонентов

1. Документирование:

-Подготовка технического описания архитектуры

-Составление руководства по использованию, справочных функций

**Анализ проблематики**

Анализ проблематики предполагает комплексное исследование ключевых аспектов разработки программного решения. В первую очередь необходимо детально изучить доступные источники информации, уделяя особое внимание их функциональным возможностям, форматам хранения и структурным особенностям организации данных о расписании. Параллельно следует проанализировать потенциальные трудности, которые могут возникнуть в процессе разработки и последующей интеграции модуля в общий проект.

Важным направлением исследования становится сравнительный анализ преимуществ и ограничений выбранного способа организации и получения данных с точки зрения его применимости в контексте создаваемого навыка. Особое внимание необходимо уделить изучению условий и технических ограничений, в рамках которых ведется разработка программного модуля.

Значительная часть анализа будет посвящена изучению существующих подходов к решению аналогичных задач, включая обзор реализованных навыков обработки расписания. При этом особый интерес представляют решения с открытым исходным кодом, позволяющие проанализировать применяемые архитектурные и алгоритмические подходы.

Во время поиска потенциальных мест, откуда возможно извлечение информации о расписании студентов и анализа источников были найдены и проработаны некоторые варианты получения необходимой информации. Также были рассмотрены источники, упомянутые в предыдущей работе. Наиболее перспективными и возможными к внедрению стали варианты самостоятельного фиксирования расписания или получения его с сайта РУЗ. К остальным источникам получения расписания можно отнести разработанный модуль по парсингу Google таблиц, расположенных на сайте расписания студентов Нижегородского кампуса НИУ ВШЭ, и извлечение данных из Единого личного кабинета студента.

В рамках прошлой курсовой работы был разработан модуль по парсингу Google Таблиц и получения данных из этого источника информации. В силу отсутствия других вариантов данный способ представлялся как практически единственно возможный, но обладающий рядом недостатков. Рассматриваемый подход обладает рядом существенных технических ограничений, требующих особого внимания. Прежде всего, система не предусматривает автоматизированного сбора данных о занятиях дополнительного профиля, включая программы Minor и курсы английского языка, поскольку эти дисциплины не входят в базовое расписание академических групп.

Существенные сложности связаны с особенностями работы с Google Таблицами как основным источником данных. Любые изменения в структуре или формате исходных таблиц требуют существенной модификации алгоритмов обработки информации. Это обусловлено жесткой зависимостью системы от конкретного способа организации и представления данных в используемом источнике.

Отдельную проблему представляет механизм авторизации через Google API. Поскольку модуль использует персональные учетные данные разработчика, периодически возникает необходимость их обновления в связи с истечением срока действия токенов доступа. Это приводит к временной потере функциональности до момента обновления аутентификационных данных.

Несмотря на все эти ограничения, в прошлой курсовой работе выбор пал на данный способ, так как он все равно позволял получать в среднем данные о расписании, которые соответствуют действительности, но в рамках данной курсовой работы по причине изменения организации хранения расписания и нахождении потенциально более универсального способа хранения и получения информации, было принято решение отказаться от использования ранее описанного подхода и полностью переписать навык.

Что касается парсинга данных с Единого личного кабинета студента, то ограничения, которые были связаны с необходимостью предоставлять конфиденциальные данные в виде пароля и логина от корпоративной почты НИУ ВШЭ остались неразрешенными, поэтому данный способ возможен только в рамках личного применения и локального запуска на устройстве пользователя.

В качестве альтернативного варианта рассматривался способ получения данных через подключение к системе РУЗ. Данный подход был признан достаточно универсальным и эффективным для разработки модуля сбора информации о расписании. В отличие от общих источников, таких как Google Таблицы, РУЗ предоставляет персонифицированные данные, включая индивидуальные занятия (например, майноры, курсы английского языка), которые отсутствуют в групповых расписаниях. Кроме того, информация в этой системе регулярно обновляется, что гарантирует её актуальность.

Применение данного метода также открывало возможность реализации функционала для преподавателей, поскольку система содержит все необходимые для этого сведения. Дополнительным преимуществом являлось наличие готовых решений в данной области, что могло сократить время разработки и минимизировать возможные ошибки за счёт анализа существующего опыта.

Однако ключевым ограничением оказался закрытый доступ к РУЗ: подключение возможно исключительно через внутренние сети НИУ ВШЭ или университетский VPN, причём только для сотрудников вуза. Однако, в рамках майнора один из участников данной курсовой работы получил доступ к файлу конфигурации VPN, что сделало возможным бы использование данного способа, но по неизвестным причинам доступ к сайту РУЗа при использовании данного файла прекратился, а с помощью заявок по получению нового файла результата достичь не удалось.

Таким образом, также, как и в прошлой работе по причине недоступности большинства способов получения расписания или ограничений, связанных с их использованием было принято решение об использовании нового подхода, а именно – сервиса YDB. Данный способ позволяет фиксировать расписание, включая английский язык, майноры и схожие предметы пользователям с различных кампусов Высшей школы экономики. Единственным ограничением является необходимость пользователю самому вносить данные о занятиях, что может быть выполнено не всегда удобно, особенно, в случае голосового управления навыком.

YDB (Yandex Database) — это распределённая отказоустойчивая СУБД, разработанная компанией Yandex для работы с большими объёмами данных и высоконагруженными сервисами. Система поддерживает горизонтальное масштабирование, что позволяет эффективно обрабатывать запросы даже при значительном росте нагрузки. YDB обеспечивает совместимость с SQL-запросами, что упрощает интеграцию с существующими решениями, а также предлагает инструменты для работы как с структурированными, так и с полуструктурированными данными (например, JSON).

Одним из ключевых преимуществ YDB является её встроенная отказоустойчивость: данные автоматически реплицируются между узлами, что минимизирует риск потери информации при сбоях. Кроме того, сервис предоставляет низкие задержки при выполнении запросов, что критически важно для голосовых помощников, где скорость ответа напрямую влияет на пользовательский опыт.

Сервис также предоставляет возможность бесплатного использования в рамках обозначенных ограничений и интегрирован в платформу Yandex Cloud, что позволяет ускорить разработку и тестирования навыка голосового ассистента. Также YDB поддерживает бессерверные вычисления в рамках работы с Yandex Cloud Functions, что избавляет от необходимости аренды, развертывания и поддержки виртуальной машины и значительно облегчает процесс выполнения работы.

Для интеграции YDB с Python используется официальный YDB Python SDK, предоставляемый Yandex Cloud. Этот SDK поддерживает все основные операции с базой данных, включая выполнение SQL-запросов, транзакционную обработку данных и работу с prepared statements для оптимизации повторяющихся запросов.

Ключевые возможности SDK:

-Подключение к YDB через gRPC-протокол, обеспечивающий высокую скорость передачи данных.

-Асинхронная работа через библиотеку asyncio, что критически важно для навыков голосовых помощников, где задержки должны быть минимальными.

-Поддержка сессионного управления, позволяющая контролировать время жизни соединения и избегать перегрузки сервера.

Данный способ хранения информации позволяет самостоятельно выбирать формат данных и сделать хранение информации более универсальным. В отличие от Google Таблиц, где данные жестко привязаны к формату ячеек, YDB позволяет хранить информацию в структурированном виде (например, в формате JSON или реляционных таблицах с четкой схемой). Это исключает проблемы, связанные с изменением оформления данных, так как система автоматически адаптируется под заданную структуру.

В Google Таблицах отсутствует возможность автоматически связывать разрозненные данные (например, название предмета с временем и аудиторией), если они не имеют строгой привязки по расположению. YDB, благодаря поддержке связей между таблицами и индексации, позволяет мгновенно находить сопутствующую информацию, даже если исходные данные меняют формат.

Также одним из главнейших преимуществ является возможность обработки расписания пользователей различных кампусов университета и корректно работать с дополнительными курсами: Minor, английский язык, различные факультативы и др. В предыдущий раз алгоритм был заложником формата и не позволял выполнять данные действия.

Хранение информации организовано в рамках реляционной базы данных[Приложение1], которая включает в себя следующие таблицы:  
1. Студент – содержит ФИО студентов, их id, id группы для связки с соответствующей таблицей.

2. Группа включает в себя id группы, название, курс, образовательную программу, факультет, формат обучения: очный, заочный или иной, уровень образования.

3. Преподаватель - ФИО, id.

4. Пара(групповая) содержит атрибуты: id, название, тип предмета(семинар, лекция), корпус, аудитория, id преподавателя, время, поля которые отражают каждую неделю проходит пара или нет и по верхним неделям или нет проходят занятия, дата и id группы. Данная таблица служит для хранения занятий, которые проводятся для групп конкретной образовательной программы.

5. Пара(индивидуальная) хранит все атрибуты групповой пары с единственным отличием, вместо id группы, хранится id студента. Соответственно, рассматриваемая таблица необходима для фиксации дисциплин, которые могут быть индивидуальны для каждого студента: английский язык, майонор, факультатив и другие.

6. Пара(групповая)-группа – вспомогательная таблица, чтобы обеспечить нормализацию данных. Необходима для связи между таблицами Группа и Пара(групповая), соответственно.

7. Преподаватель-группа – вспомогательная таблица.

8. Студент-пара(индивидуальная) – вспомогательная таблица.

Подводя итог, можно сказать, что YDB является оптимальным решением для осуществления организации хранения, обработки и получения информации о пользователях и их расписании и обладает рядом преимуществ по сравнению с рассмотренными выше способами: универсальностью, гибкостью, быстрой и эффективной возможностью выполнять поиск против одного недостатка - необходимости самостоятельно заполнять данные о расписании.

В рамках исследуемой проблематики было проанализировано несколько существующих решений, однако ни одно из них не может быть использовано в качестве полноценной альтернативы.

Первое рассмотренное решение требует исключительно ввода кода, полученного от бота, не предоставляя пользователю никаких дополнительных сведений или инструкций. Второй вариант – "Помощник Вышки" – оказался неработоспособным: при обращении к нему система выдает сообщение об ошибке, указывающее на невозможность получения ответа. Хотя исходный код данного навыка доступен на GitHub, его функциональность ограничена из-за отсутствия доступа к API Расписания учебных занятий (РУЗ), что приводит к ошибкам при попытке запуска. Таким образом, на текущий момент в открытых источниках не обнаружено готового решения, соответствующего поставленным в проекте задачам. Выполненный в прошлом году программный модуль на данный момент нефункционален в силу изменения способа хранения информации и не обладает рядом существенных недостатков по сравнению с новым выбранным вариантом реализации.

Что касается опыта других учебных заведений, то было выявлено два подхода. Финансовый университет разработал специализированное приложение, интегрированное с внутренним API расписания. Аналогичное решение существует в Томском политехническом университете, где расписание доступно через отдельный веб-сайт. Прочие аналогичные системы, функционирующие в данной предметной области, используют более совершенные и универсальные методы получения данных, что обеспечивает их стабильную работу.

Данный анализ подтверждает актуальность разработки нового решения, поскольку существующие аналоги либо неработоспособны, либо адаптированы под специфические условия других образовательных учреждений.

**Анализ существующих решений, технологий и аналогов**

В рамках рассматриваемой предметной области существует несколько различных подходов, позволяющих реализовать получение необходимой информации из Google Таблиц. Так как информация о расписании студентов Высшей школы экономики располагается на сайте, то существует некоторое количество вариантов его извлечения, которые предоставляют возможность использовать различные языки программирования, таким образом общее число вариантов реализации технического решения вырастает многократно. Первый способ – использование существующих технологий и библиотек для получения данных с самого сайта, на котором располагается таблица с расписанием. Для данной задачи подходят различные языки программирования, такие как JavaScript, Python, PHP, C++ и другие, которые имеют разработанные библиотеки для веб-скрапинга. Второй способ – использование API Google Таблиц для работы уже с самими таблицами и обработки и извлечения данных из них. Рассматриваемое техническое решение также может быть также реализовано с применением широкого спектра языков программирования, таких как Python, Go, JavaScript, Java, Google Apps Script.

При получении информации с сайта используемый программный модуль отправляет HTTP-запросы к серверу, требуя то содержимое, которое расположено на самой странице. Существует два вида сайтов статические и динамические.

Первые состоят из постоянных страниц, которые не меняют содержимое в зависимости от пользователя. При выполнении запроса сервер сразу предоставит готовый HTML код в том формате, который был использован при его создании.

Динамические сайты представляют собой страницы, которые в зависимости от пользователя будут иметь различную организацию данных. Данный тип страниц не существует в исходном виде на сервере, а при каждом новом запросе они формируются. Сначала сервер находит нужный документ и отправляет его интерпретатору, который выполняет код из HTML-документа и сверяется с файлами и базой данных. После этого документ возвращается на сервер и затем отображается в браузере. Для интерпретации страниц на серверной стороне используются языки программирования Java, PHP, ASP и другие. [4]

В зависимости от типа сайта необходимо использование библиотек и технических решений, способных выполнять отправку запроса на сервер и получение нужного кода HTML для его дальнейшей обработки. Так, например, в языке Python существуют библиотеки requests и selenium, которые предназначены для парсинга разного формата данных, первая подходит только для извлечения данных из статических сайтов, вторая, в свою очередь, годится для получения информации в независимости от типа сайта.

В случае рассматриваемой проблематики, информационное решение должно уметь отправлять запросы и получать информацию со статического сайта, так как таблица, которая располагается на нем не изменяется и одинакова для любого пользователя. Для решения поставленной задачи, как было упомянуто ранее возможно использование нескольких языков программирования.

JavaScript – является одним из наиболее распространенных решений, подходящих для выполнения широкого круга задач. Это обосновано прежде всего наличием значительного количества библиотек и инструментов, которые доступны для решения задач получения информации с сайта и поддержкой со стороны сообщества разработчиков и энтузиастов. Также в пользу данного языка говорит простота синтаксиса и наличие большого спектра ресурсов для обучения. К числу библиотек, которые подходят для решения существующей задачи подходят Unirest, Axios и Cheerio, которые не используют безголовые браузеры и драйверы и могут быть использованы в рамках Yandex Functoins. Все рассматриваемые решения поддерживают различные HTTP-методы, такие как GET, POST и другие и позволяют выполнять основную задачу, а именно: получение HTML кода страницы с расписанием.

С++ - язык программирования, который является наиболее высокоуровневым по сравнению с JavaScript и Python. Данный факт дает ему преимущество в плане гибкости и скорости, которое следует за сложностью синтаксиса и концепции языка, а также низкой скоростью разработки. Способность получать доступ к низкоуровневым системным ресурсам дает возможность создавать настраиваемые решения для парсинга. Также C++ является компилируемым языком, что предоставляет ему преимущество по сравнению с JavaScript и Python, которые требуют наличие интерпретатора, что влечет большое использование ресурсов и более низкую производительность. К основным библиотекам, которые используются для парсинга сайтов относятся libcurl, Boost.Asio, htmlcxx. Данные решения также позволяют получать HTML-содержимое необходимого сайта и обрабатывать его для извлечения нужного текста.

Python, в свою очередь, предлагает значительное количество библиотек для веб-скрапинга, позволяющие организовать технические решения, которые будут уникальны для каждого проекта. Также одними из основных преимуществ данного языка являются простой синтаксис и активная поддержка со стороны энтузиастов.

Учитывая контекст рассматриваемой проблемы, а именно: использование другими участниками Python в качестве языка программирования, что позволяет удобно объединять разрабатываемые подпрограммы в проект, отсутствие необходимости в анализе большого количества данных, где была бы нужна производительность языка C++, наличие опыта разработки на Python, для реализации программного модуля по извлечению расписания был выбран данный язык.

Для решения поставленной задачи выбранное техническое решение предполагает использование таких библиотек как Requests, Beautiful Soup, и Selenium. Последняя имеет возможность парсинга динамических веб-сайтов, что не имеет необходимости при парсинге рассматриваемого сайта с таблицей. Кроме того, использование Selenium влечет необходимость установки соответствующего драйвера браузера, который запускается и используется данной библиотекой. Для выполнения этих условий в среде запуска необходимо присутствие виртуальной машины, которая бы запускала драйвер, что усложняет сам процесс разработки, также не было найдено решения для этой проблемы в Yandex Cloud Functions. Библиотека Requests, в свою очередь, предоставляет возможность парсинга веб-страниц буквально в пару строк. Необходимо отправить GET-запрос на необходимый сайт с помощью функции get и с применением метода text получить HTML-код страницы. Также отсутствует необходимость использования других решений, кроме самой библиотеки. Анализ же самого кода выполняется с помощью библиотеки Beautiful Soup, которая позволяет получать тексты, атрибуты, ссылки и другие данные из документа на основе тегов, классов, идентификаторов, атрибутов и других характеристик элементов. Например, вы можете найти все ссылки или изображения на странице с определенным классом или атрибутом. Однако в существующем контексте, когда необходимо получение всей информации и при условии, что не было найдено возможных закономерностей по названию имен классов, тегов или других элементов текстов, которые бы позволили каким-либо образом структурировать информацию, было принято решение использования и поиска другого технического средства. Пример получаемой информации: </span> НИС &quot;Анализ литературы&quot; - Савченко Л.В.</td><td class="s8" dir="ltr">301<br><br>303<br><br>301</td><td.

Помимо вышеупомянутой технологии также существует другой способ взаимодействия с информацией, а именно: применение API Google Таблиц, который поддерживает различные языки программирования такие как Python, Go, JavaScript, Node.js, Java и другие. По ранее указанным причинам, для работы с изучаемым техническим решением был выбран Python. Касаемо языка программирования, необходима версия Python, начиная с 3.10.7 и новее, также нужен модуль, который организовывает управление пакетами pip, имеющийся проект в Google Cloud Platform.

При работе с API Google таблиц необходимо определенная настройка своего проекта в Google Cloud Platform, откуда будут загружены учетные данные разработчика для авторизации при работе с таблицами. Выбранный способ реализации подпрограммы требует наличия клиентских библиотек Google для работы с Python: Google-api-python-client, Google-auth-httplib2, Google-auth-oauthlib.

Для работы с Google Таблицами также возможно использование других библиотек, таких как Gspread или Pygsheets, однако в контексте рассматриваемой проблемы, когда необходимо только получение значений из таблицы, данные модули являются эквивалентными, так как получение информации происходит в рамках одного метода и основной задачей является уже работа с полученными сведениями.

Для работы программы необходима настройка среды исполнения API Google Таблиц, которая подразумевает несколько шагов, указанный в Кратком руководстве по Python.[5] Прежде чем использовать API Google, необходимо включить их в проекте Google Cloud. Можно задействовать один или несколько API в одном проекте Google Cloud. Затем происходит настройка экрана согласия OAuth. Если используется новый проект Google Cloud для выполнения краткого руководства, то необходимо настроить экран согласия OAuth и добавить собственный аккаунт в качестве тестового пользователя. Для этого нужно:

1. В консоли Google Cloud выбрать >  **API и службы** > **Экран согласия OAuth**.
2. В качестве **типа пользователя** выбрать **«Внутренний»**, затем нажать **«Создать»** .
3. Заполнить форму регистрации приложения, затем нажать **«Сохранить и продолжить»**.
4. На данный момент можно пропустить добавление областей и нажать **«Сохранить и продолжить»**. В будущем, когда будет необходимость создания приложение для использования за пределами организации Google Workspace, необходимо изменить **тип пользователя** на **Внешний**, а затем добавить области авторизации, необходимые приложению.
5. Просмотреть сводку регистрации приложения. Чтобы внести изменения, нажать **«Изменить»**. Если регистрация приложения выглядит нормально, нажать **«Вернуться на панель управления»**.

Чтобы аутентифицировать конечных пользователей и получить доступ к пользовательским данным в приложении, необходимо создать один или несколько идентификаторов клиентов OAuth 2.0. Идентификатор клиента используется для идентификации одного приложения на серверах Google OAuth. Если приложение работает на нескольких платформах, необходимо создать отдельный идентификатор клиента для каждой платформы.

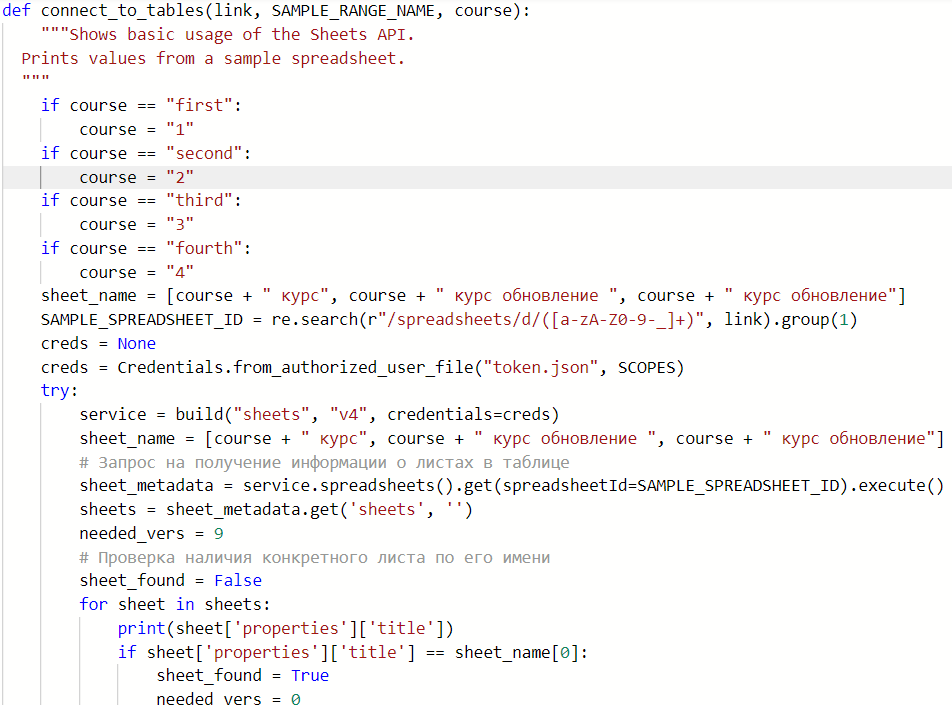
1. В консоли Google Cloud выбрать  > **API и службы** > **Учетные данные** .
2. Нажать **«Создать учетные данные»** > **«Идентификатор клиента OAuth** ».
3. Нажать **Тип приложения** > **Приложение для ПК** .
4. В поле **Имя** ввести имя учетных данных. Это имя отображается только в консоли Google Cloud.
5. Нажать **Создать** . Появится экран создания клиента OAuth, показывающий новый идентификатор клиента и секрет клиента.
6. Нажать **ОК** . Вновь созданные учетные данные появятся в разделе **«Идентификаторы клиентов OAuth 2.0».**
7. Сохранить загруженный файл JSON как credentials.json и переместить его в свой рабочий каталог.

Таким образом, для реализации программного модуля был выбран язык Python за счет отсутствия необходимости в высокой производительности, применение данного решения другими участниками проекта, что упрощает процесс интеграции программного модуля в совокупности с клиентскими библиотеками Google, так как при решении поставленной задачи существенной разницы между ней и аналогами не было обнаружено.

**Полученные результаты, методика испытаний и итоги анализа**

В процессе решения поставленной задачи был реализован программный модуль, состоящий из нескольких функций и обрабатывающий необходимую Google Таблицу с получением массива нужного расписания для всех групп образовательной программы. Для его работы требуется выполнение, помимо наличия указанных клиентских библиотек Google и файла с учетными данными пользователя, выполнение шагов по настройке среды, необходимо также импортирование библиотеки Re, которая будет использована для отделения из всех поступающих данных атрибутов занятия, а именно: времени, лектора, аудитории и типа занятия.

Первая функция, задействованная в программном модуле, называется connect\_to\_tables. На вход принимается переменная link, полученная, как результат функции take\_a\_link из подпрограммы другого участника проекта и представляющая собой ссылку на сайт с Google Таблицей расписания образовательной программы пользователя. Помимо этого, при запуске рассматриваемой функции необходимо передать в качестве аргумента переменную SAMPLE\_RANGE\_NAME - диапазон ячеек, который необходимо получить. Значения данной переменной отличается для очных и очно-заочных образовательных программ из-за того, что таблицы представлены для каждого вида обучения в различном формате и является фиксированным для всех образовательных программ каждого вида. Также на вход принимается переменная course, отвечающая за курс пользователя и необходимая для парсинга нужного листа, так как расписание всех курсов образовательной программы находятся в одной таблице.(рисунок 1)



Из полученной ссылки с помощью регулярного выражения делается срез строки с необходимым идентификационным номером таблицы, который требуется для получения всех данных. Для работы данной функции также необходимо наличие учетных данных пользователя в файле token.json. В изначальной версии функции, которая взята с сайта[2] в случае, если они отсутствуют, то запускаются методы, которые авторизуют пользователя через браузер и создают нужный файл. Однако среда Yandex Cloud Functions не поддерживает запись файлов, поэтому данная часть кода была удалена, но возникла необходимость периодически, в промежуток около 4-5 дней обновлять token.json вручную и загружать его на используемую платформу, так как учетные данные, записанные в данном файле, устаревают и возникает ошибка, не позволяющая продолжить работу функции.

После создания переменной с учетными данными вызываются методы библиотек Google, осуществляющие парсинг всей информации из таблицы нужного листа, которая возвращается функцией в виде переменной values.

В случае, если не будут найденные данные или не найден соответствующий лист вызываются исключения NoValuesExeption, No\_such\_sheet\_exeption, наследующиеся из класса Exception.

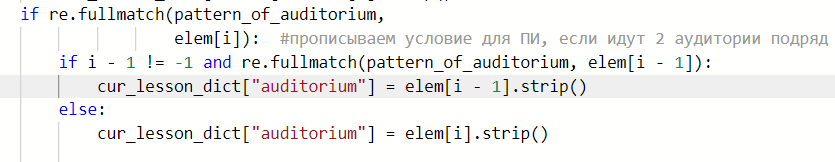
Данные, полученные в ходе скрапинга листа таблицы представляют собой двумерный вложенный список. Именно данный список обрабатывается функцией table\_parsing. Эта часть программы принимает в качестве аргументов link – ссылка на расписание образовательной программы, course – номер курса, которые необходимы для вызова функции connect\_to\_tables.

В рамках данной подпрограммы происходит обработка имеющейся информации о студентах. Сначала с помощью регулярный выражений выделяются названия групп, которые располагаются во втором вложенном списке переменной values. Затем создается список словарей schedule\_res, в которых ключом выступает название группы, а значением – вложенный словарь, где день недели – ключ, а значение еще один вложенный словарь, где уже идентификатором является строка “lessons”, по которой можно получить список словарей, в которых указаны необходимые атрибуты занятия: "name", "time", "auditorium", "lecturer", "lesson\_type". В данном виде идет организация всей информации, полученной в результате обработки переменной values.(рисунок2) 

После создания пустого макета происходит замена у каждого элемента вложенного списка values символов переносов строки и дефисов на пробелы, потому что, в зависимости от таблицы, между элементами переменной с ее содержанием может содержаться разное число данных знаков, которые для удобства изменяются на пробел.

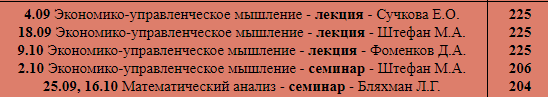
Далее происходит уже получение необходимых элементов расписания за счет выявленной в таблице закономерности, под которой подразумевается тот формат, в котором хранятся и передаются элементы вложенных списков переменной values. А именно: если ячейка хранит информацию о расписании, то место первого элемента списка зарезервировано под название дня недели, затем идет время, пустая строка, название предмета, аудитория, пустая строка, название предмета для следующей группы, аудитория и так далее. В случае, если занятий в данное время и день недели нет, то на месте имен предметов и аудиторий идут пустые строки. Также день недели обозначается только во вложенном списке с временем 8:00-9:20.

Используя выявленную закономерность, идет итерирование по всем элементам вложенных списков исходной переменной. Для определения, что элемент является временем или аудиторией используется регулярное выражение. Для каждой группы и дня недели заполняются словари cur\_lesson\_dict(рисунок 2). Инкрементация дня недели происходит, если был обнаружен элемент, который ему соответствует, а группы, если текущее число пустых строк превысило 3 или был найден элемент, который является аудиторией. Обнуление номера группы выполняется, в свою очередь, в случае перехода к новому вложенному списку. При тестировании было обнаружено, что в расписании 1 курса Программной инженерии в таблице встречается элемент, который вместо пустой строки представляет собой аудиторию, визуально не обозначенную в таблице. Для корректности работы было прописано условие(рисунок 3), во время возникновения которого в качестве аудитории берется предыдущий элемент.



Функция table\_parsing возвращает кортеж, состоящий из расписания в указанном формате с пустыми значениями для "lecturer", "lesson\_type", так как на данном этапе в связи с различным форматом таблиц их определить сложно, и список с названиями учебных групп.

С расписанием в описанном формате далее происходит работа с помощью функции lessons\_split, решающей проблемы, связанные с некоторыми особенностями записи информации в таблицах. Входными данными является та информация, которая возвращается предыдущей функцией. Первый шаг – работа с ячейками, в которых указано несколько предметов и аудиторий. Например, в расписании группы 23МББЭ1 (рисунок 4), данные приходят в формате, где в одной ячейке содержится информация о всех предметах.

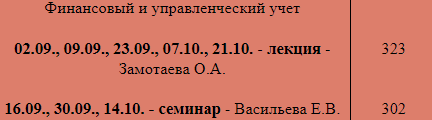


Первый вложенный цикл обрабатывает каждый элемент списка словаря по ключу “lessons”. В случае, если обнаруживается, что элемент расписания auditorium можно разбить на несколько элементов, то создаются списки, содержащие название, лекторов и аудитории каждых предметов. Производится итерирование по каждому элементу данных списков и заполняется переменная cur\_lesson\_dict, аналогичная той, что содержится в функции table\_parsing, затем она добавляется в исходный список schedule\_res, а элемент, содержащий информацию о нескольких занятиях удаляется.

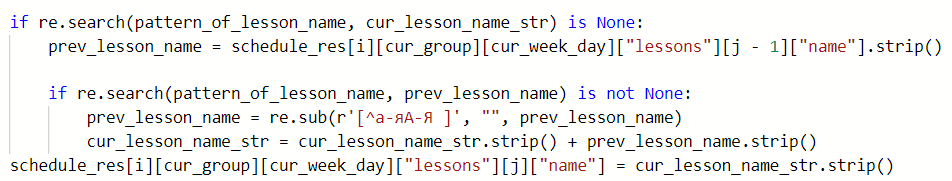
На данном этапе также происходит определение и заполнение параметра занятия lecturer. Для определения данного элемента пары используется регулярное выражение, которое учитывает варианты, когда преподаватель обозначается несколькими способами – комбинацией фамилии, инициалов и точек. Возникает проблема, когда лектор не обозначен или обозначен только фамилией, в этом случае подобрать необходимое регулярное выражение не удалось, так как оно соответствовало бы названию предмета.

Другой особенностью формата является наличие одной аудитории для двух занятий. Данный случай обрабатывается в следующем цикле. Идет работа также с элементом name каждого занятия и в случае, если при разбиении по паттерну преподавателя элемент становится списком длиной более 1, то идет итерация по полученному списку и занесение информации в исходный список всех групп. В исходной ячейке с несколькими предметами все значения приравниваются к пустой строке.

Затем идет обработка оставшихся предметов, которые не должны иметь никаких особенностей, связанных с наличием более одного предмета и аудитории в соответствующих значениях словаря. С применением регулярных выражений идет также итерация по каждому предмету и поиск паттернов, которые соответствуют лекторам и типам предметов. В данную часть подпрограммы также включена обработка ситуации. (рисунок 5)



В данном случае название предмета будет хранится только под соответствующим ключом предыдущего предмета. Если было обнаружено, что в name текущего занятия нет элемента, удовлетворяющего соответствующему паттерну, то выбирается часть названия предыдущего предмета и конкатенируется с соответствующей строкой текущего.(рисунок 6)



После данной обработки рассматриваемая функция возвращает список с полученными данными в указанном выше формате. Данный формат используется только в рамках тестирования и отладки подпрограммы. Для удобства работы с информацией он преобразуется в формат списка с объектами классов Group.

Последняя функция, которая присутствует в подпрограмме по парсингу расписания - transform\_to\_classes. На вход подается переменная schedule\_res. Происходит итерация по каждому элементу данного объекта с заполнением атрибутов объекта Group, которыми являются: название группы – поле класса string, день недели – list. В день недели, в свою очередь попадают объекты класса Lesson с заполненными атрибутами time, lecturer, name, lesson\_type и auditorium. Данный формат является итоговым выводом программного модуля, затем идет обработка полученной информации и нахождение необходимых элементов.

В случае возникновения ошибок в коде, которые в силу отличий формата заполнения таблиц могут появляться в различных местах программы, блок каждой функции помещен в конструкцию try – exception, при возникновении ошибки она печатается в консоли, а пользователю выводится сообщение: "Oшибка доступа к серверу, давайте я расскажу вам последние новости".

Для тестирования функции были выбраны расписания различных образовательных программ и курсов. Проверка работоспособности проводилась в рамках IDE PyCharm.

Сначала проверялось время работы подпрограммы. Для этого были использованы таблицы с расписанием во втором модуле. Время колебалось от 1 до примерно 2.6 секунд, что в контексте, когда ответ навыка ограничен 4.5 секундами представляется приемлемым.

Помимо этого, была протестирован факт запуска без ошибок программного модуля на всех образовательных программах Нижнего Новгорода в 1 и 2 модулях, было выявлено, что алгоритм выполняется на всех направлениях и курсах кроме Управления бизнесом в силу особенностей заполнения таблиц.

Также была проверена корректность отображения элементов расписания на выборке во втором модуле(таблица 1)

|  |  |  |
| --- | --- | --- |
| **Данные** | **Корректность отображения** | **Неучтенные элементы** |
| ОП - Юриспруденция Группы - 23Ю | Отделение нескольких предметов и одной аудитории корректно и все также | Не учтены случаи, когда лекция и семинар в одной клетке, некорректное отображение типа предмета |
| ОП - Математика Группы - 21ФМ | Абсолютно корректно | Нет |
| ОП - Прикладная математика и информатика  Группы - 22ПМИ | Абсолютно корректно | Нет |
| ОП - Бизнес-информатика  Группы - 22БИ | В целом корректно | Некорректное отделение 30.11 МБП как отдельного предмета, в параметр lecturer записан только один лектор |
| ОП - Компьютерные науки и технологии  Группы - 23КНТ | Абсолютно корректно | Нет |
| ОП - Цифровой маркетинг  Группы - 23ЦМ | Некорректно | Предметы во вторник в одной аудитории не отделяются, отделение слова отмена в отдельный предмет |
| ОП - Экономика | Некорректно | Отсутствие разделения предметов в одной аудитории, лишнее разделение, некоторые занятия не учтены |

**Список использованных источников**

1. Рынок разговорного AI в России 2020-2025[Электронный ресурс].: официальный сайт- <https://just-ai.com/blog/issledovanie-rynok-razgovornogo-ii-v-rossii-2020-2025> (дата обращения: 16.01.2024)
2. Сайт с расписанием Высшей школы Экономики Нижнего Новгорода[Электронный ресурс].: официальный сайт - <https://nnov.hse.ru/uch/schedule> (дата обращения: 19.01.2024)
3. Обзор API Google Таблиц: официальный сайт[Электронный ресурс]. - <https://developers.google.com/sheets/api/guides/concepts?hl=ru> (дата обращения: 20.02.2024)
4. Статические и динамические сайты сегодня: какие лучше и почему[Электронный ресурс]. — URL: https://o.jino.ru/journal/articles/staticheskie-dinamicheskie-sayty/ (дата обращения: 23.02.2024).
5. Краткое руководство по Python [Электронный ресурс]. — URL: <https://developers.google.com/sheets/api/quickstart/python?hl=ru> (дата обращения: 03.03.2024)
6. Документация языка Python: [Электронный ресурс]. официальный сайт- <https://www.python.org/doc/> (дата обращения: 04.03.2024)
7. Документация Yandex Cloud: [Электронный ресурс] официальный сайт.т - <https://yandex.cloud/ru/docs/functions/tutorials/alice-skill?utm_referrer=https%3A%2F%2Fyandex.ru%2F> (дата обращения: 05.03.2024)
8. Школа Алисы: официальный сайт[Электронный ресурс]. — URL: https://events.yandex.ru/events/webinars/alisa-school/index (дата обращения: 06.03.2024)