Муниципальное бюджетное общеобразовательное учреждение

средняя общеобразовательная школа №21

Индивидуальный проект

Тема: Парсинг сайта на Python с отправкой сообщений в Telegram бота

Тип проекта: исследовательский

Допущена к защите « » 2022 г .

Подпись рук-ля

Выполнила: Пунич Полина, 11 А класс

Канск

2022 г.

**СОДЕРЖАНИЕ**

|  |  |
| --- | --- |
| Тема проекта | Парсинг сайта на Python с отправкой сообщений в Telegram бота на aiogram |
| Образовательное учреждение | МБОУ СОШ №21 |
| Автор проекта | Пунич Полина 11 а |
| Целевая группа | Подростки, студенты |
| Сроки реализации проекта | апрель |
| Тип проекта | WEB |
| Партнёры |  |
| Ожидаемые результаты | Бот в telegram будет успешно выводить товары, опубликованные на сайте |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Этап | Действия | Сроки | Результат | Участники |
| Подготовительный | Поиск идеи проекта | Октябрь 2021 | Определилась тема проекта, проблема, цель | Пунич П.П. |
| Основной, аналитический | Поиск информации | Апрель-май | Систематизирование информации, ее анализ, отработка | Пунич П.П. |
| практический | Работа над идеей | Июнь- сентябрь |  | Пунич П.П. |
| Работа над папкой проекта | Январь- |  | Пунич П.П.  Шуляк Н.В. |
| Заключительный |  |  |  |  |

ВВЕДЕНИЕ

**Актуальность:** среди подростков и взрослых почему?????

**Проблема:** пользователь редко посещает ресурсы/сайты для просмотра новостей или сайт недоступен в других странах

**Цели**: пользователь смог рассматривать новости сайта через telegram

**Гипотеза:** программа будет выводить новые новости, опубликованные на сайте

**Объект исследования:** Lenta.ru

**Предмет исследования:** Python, Telegram

**Задачи:**

1. Рассказать о языке программирования Python
2. Рассказать о библиотеках, которые понадобятся для программы
3. Изучить способы парсинга сайтов
4. Реализовать бота
5. **Python**
   1. **Общие факты**

**Python** — [высокоуровневый язык программирования](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D1%81%D0%BE%D0%BA%D0%BE%D1%83%D1%80%D0%BE%D0%B2%D0%BD%D0%B5%D0%B2%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) общего назначения с [динамической](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D0%BD%D0%B0%D0%BC%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%82%D0%B8%D0%BF%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F) [строгой](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D0%BE%D0%B3%D0%B0%D1%8F_%D1%82%D0%B8%D0%BF%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F) типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости [кода](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D1%85%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D0%B4) и его качества, а также на обеспечение переносимости написанных на нём программ. Язык является полностью [объектно-ориентированным](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%8A%D0%B5%D0%BA%D1%82%D0%BD%D0%BE-%D0%BE%D1%80%D0%B8%D0%B5%D0%BD%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) в том плане, что всё является [объектами](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%8A%D0%B5%D0%BA%D1%82_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)). Необычной особенностью языка является выделение [блоков кода](https://ru.wikipedia.org/wiki/%D0%91%D0%BB%D0%BE%D0%BA_%D0%BA%D0%BE%D0%B4%D0%B0) пробельными отступами. [Синтаксис](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%81_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) ядра языка минималистичен, за счёт чего на практике редко возникает необходимость обращаться к документации. Сам же язык известен как [интерпретируемый](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D1%80%D0%BF%D1%80%D0%B5%D1%82%D0%B8%D1%80%D1%83%D0%B5%D0%BC%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) и используется в том числе для написания [скриптов](https://ru.wikipedia.org/wiki/%D0%A1%D0%BA%D1%80%D0%B8%D0%BF%D1%82).

**Python** является [мультипарадигмальным языком программирования](https://ru.wikipedia.org/wiki/%D0%9C%D1%83%D0%BB%D1%8C%D1%82%D0%B8%D0%BF%D0%B0%D1%80%D0%B0%D0%B4%D0%B8%D0%B3%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F), поддерживающим [императивное](https://ru.wikipedia.org/wiki/%D0%98%D0%BC%D0%BF%D0%B5%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5), [процедурное](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D1%86%D0%B5%D0%B4%D1%83%D1%80%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5), [структурное](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5), [объектно-ориентированное](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%8A%D0%B5%D0%BA%D1%82%D0%BD%D0%BE-%D0%BE%D1%80%D0%B8%D0%B5%D0%BD%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) программирование, [мета программирование](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%B0%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) и [функциональное](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) программирование.

[Стандартная библиотека](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B0_Python&action=edit&redlink=1) включает большой набор полезных переносимых функций, начиная с возможностей для работы с текстом и заканчивая средствами для написания сетевых приложений. Дополнительные возможности, такие как математическое моделирование, работа с оборудованием, написание веб-приложений или разработка игр, могут реализовываться посредством обширного количества сторонних библиотек, а также интеграцией библиотек, написанных на Си или C++, при этом и сам интерпретатор Python может интегрироваться в проекты, написанные на этих языках. Существует и специализированный [репозиторий](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BF%D0%BE%D0%B7%D0%B8%D1%82%D0%BE%D1%80%D0%B8%D0%B9) программного обеспечения, написанного на Python, — [PyPI](https://ru.wikipedia.org/wiki/PyPI). Данный репозиторий предоставляет средства для простой установки пакетов в операционную систему и стал стандартом де-факто для Python. По состоянию на 2019 год в нём содержалось более 175 тысяч пакетов.

Python стал одним из самых популярных языков, он используется в [анализе данных](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), [машинном обучении](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), [DevOps](https://ru.wikipedia.org/wiki/DevOps) и [веб-разработке](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%B1-%D1%80%D0%B0%D0%B7%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0), а также в других сферах, включая [разработку игр](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B7%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D1%85_%D0%B8%D0%B3%D1%80). За счёт читабельности, простого синтаксиса и отсутствия необходимости в компиляции язык хорошо подходит для обучения программированию, позволяя концентрироваться на изучении алгоритмов, концептов и парадигм. Отладка же и экспериментирование в значительной степени облегчаются тем фактом, что язык является интерпретируемым. Применяется язык многими крупными компаниями, такими как [Google](https://ru.wikipedia.org/wiki/Google_(%D0%BA%D0%BE%D0%BC%D0%BF%D0%B0%D0%BD%D0%B8%D1%8F)) или [Facebook](https://ru.wikipedia.org/wiki/Facebook)[[25]](https://ru.wikipedia.org/wiki/Python#cite_note-:2-25). По состоянию на октябрь 2021 года Python занимает первое место в [рейтинге TIOBE](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%B4%D0%B5%D0%BA%D1%81_TIOBE) популярности языков программирования с показателем 11,27%. «Языком года» по версии TIOBE Python объявлялся в 2007, 2010, 2018 и 2020 годах.

1. **Новостные ленты**

**Новостная лента** — формат данных, используемый для доставки пользователям часто обновляемой информации. Распространители этой информации предоставляют новостную ленту, позволяя пользователям подписаться на неё.

Лента состоит из некоторого ограниченного числа статей, а также из служебной информации типа автора ленты или домашней странички. При появлении новых данных они добавляются в ленту, вытесняя при этом старые статьи. Обычно в ленте бывает не больше 10-20 статей, хотя бывают ленты и с несколькими сотнями статей.

Новостные ленты автоматически собираются новостными агрегаторами, которые могут быть либо веб-приложениями, либо устанавливаться на компьютер конечного пользователя.

Типичный сценарий использования новостной ленты таков: создатель контента публикует ссылку на ленту на своём сайте, которую конечный пользователь может добавить в свой агрегатор (также известный как клиент для чтения лент новостей), который работает на компьютере конечного пользователя. Через заданные пользователем интервалы, либо по требованию агрегатор скачивает новые версии лент новостей и, если в них появились новые записи, каким-либо способом уведомляет пользователя.

При помощи лент новостей может доставляться обычный HTML-контент (веб-страницы), ссылки либо другие типы цифровых данных. Например, в статьях в ленте, посвящённой какому-нибудь сериалу или подкасту может присутствовать особым образом сформированная ссылка, при обнаружении которой поддерживающие эту технологию агрегаторы скачают соответствующий файл. Таким образом реализуется периодическая доставка цифровых медиафайлов, также известная как Broadcatching.

Новостные ленты предоставляются многими новостными сайтами и блогами.

1. **Telegram**
   1. **Telegram — это…**

Telegram — кроссплатформенная система мгновенного обмена сообщениями (мессенджер) с функциями VoIP, позволяющая обмениваться текстовыми, голосовыми и видеосообщениями, стикерами и фотографиями, файлами многих форматов. Также можно совершать видео- и аудиозвонки и трансляции в каналах и группах, организовывать конференции, многопользовательские группы и каналы. С помощью ботов функционал приложения практически не ограничен. Клиентские приложения Telegram доступны для Android, iOS, Windows, macOS и GNU/Linux. В июне 2022 года вошёл в пятёрку самых загружаемых приложений, а число его постоянных пользователей превысило 700 миллионов.

По данным исследования TGStat на май 2019 года, наибольшая доля пользователей из России приходится на лиц в возрасте 25—34 (38 %) и 18—24 (27 %) лет. 35,6 % российских пользователей составляют жители Москвы, 13,3 % — жители Санкт-Петербурга. Остальные регионы распределены более равномерно. Исследование проводилось с помощью ответов примерно 82 тысяч респондентов разного пола, возраста и местоположения.

Проект финансируется Павлом Дуровым в объёме порядка 13 млн долларов США ежегодно.

Помимо обмена сообщениями в диалогах и группах, в мессенджере можно хранить неограниченное количество файлов, вести каналы (микроблоги), создавать и использовать ботов.

C 16 апреля 2018 года по 18 июня 2020 года на использование мессенджера на территории России были наложены ограничения.

|  |
| --- |
| **Особенности** |
| **Telegram**— [облачный](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D0%BB%D0%B0%D1%87%D0%BD%D0%BE%D0%B5_%D1%85%D1%80%D0%B0%D0%BD%D0%B8%D0%BB%D0%B8%D1%89%D0%B5) мессенджер, его можно использовать одновременно на нескольких устройствах, и все чаты и файлы (за исключением секретных чатов) будут доступны на этих устройствах. Во всех чатах можно использовать голосовые сообщения, видеосообщения, прикрепление файлов, стикеры, gif-анимации и [эмодзи](https://ru.wikipedia.org/wiki/%D0%AD%D0%BC%D0%BE%D0%B4%D0%B7%D0%B8); есть отметка о том, что собеседник прочитал сообщение и т. д. |

* 1. **Телеграм-боты — это…**

Также через мессенджер можно узнавать новости, быть в курсе валют, переводить тексты, играть в игры и еще много всего. А выполнять все эти действия помогают **специальные роботы – Телеграм-боты**.

Название «бот» происходит от сокращенного слова «робот», чем он и является. **Телеграм-бот** – это специальный аккаунт, созданный в автоматическим режиме, который позволяет пользователям совершать разные действия через сам мессенджер.

Роботы выполняют действия по текстовым командам пользователя, по принципу «вопрос-ответ» после нажатия кнопки «Старт».

Робота определить легко, в его названии всегда есть слово «bot», а еще он не может сам первый начать беседу и не имеет статуса «онлайн» или «не в сети» – вы всегда будете видеть подпись «bot».

Сегодня роботы стали очень популярны, они помогают пользователям выполнять типичные рутинные действия в автоматическом режиме, значительно упрощая им жизнь. Для владельцев же самих каналов боты стали незаменимыми помощниками в работе.

**Телеграм-боты имеют множество очевидных плюсов:**

* Доступны 24/7;
* Моментальный ответ пользователю;
* Удобство в пользовании, общение по принципу «вопрос-ответ» и текстовые задания под силу давать даже совсем неопытному пользователю мессенджера;
* Не требуют установки дополнительных программ, приложений и т.п. Все общение с ботом ведется напрямую через мессенджер;
* Безопасность личных данных – боты работают исключительно по заданным командам.
* Неограниченные возможности – виртуального помощника можно запрограммировать на отправку новостей, рассказ анекдотов, напоминание важной информации, поиск заведений, бронирование столиков в ресторане, заказ билетов и т.д.

**Функции Телеграм-бота**

Как говорят в Телеграм: «Он только посуду мыть не умеет». Действительно боты выполняют множество функций, и вот лишь небольшой их перечень:

1. **Развлечение**. Боты могут присылать смешные мемы, картинки, анекдоты, помогают выбрать фильм, найти песню по голосовому сообщению и не только.
2. **Поиск и обмен файлов**. Бот помогает отправлять и сохранять файлы с разных источников, находить Торренты, электронные книги и пр.
3. **Новости, важная информация**. Бот осветит новости, погоду, курсы валют и т.д.
4. **Утилиты и инструменты**. Робот помогает переводить тексты, напоминать о важных событиях и т.д.
5. **Интеграция с другими сервисами**. Робот может отправлять уведомления, управлять умным домом и т.д.
6. **Поиск мест**. Телеграм-бот помогает искать гостиницы, кинотеатры, рестораны и другие заведения.
7. **Транзакции**. Робот позволяет бронировать билеты, делать заказы, вызывать такси и пр.
8. **Парсинг**

**Что такое парсинг**

**Парсинг** — автоматизированный сбор и систематизация информации из открытых источников с помощью скриптов. Другое название этого процесса — веб-скрейпинг.

Скрипты, которые собирают и систематизируют информацию, называются парсерами. Они работают так:

* ищут источники по заданным параметрам — например, парсеру можно дать список сайтов, а он найдёт на них страницы с ценами;
* извлекают из источников нужную информацию — несколько строк текста, ссылку или сумму;
* преобразуют информацию — например, парсер может взять фрагмент из HTML-документа и преобразовать его в текст без кода;
* сохраняют информацию в нужном формате — например, в виде списка или таблицы в Excel.

Парсеры работают на разных языках программирования — Python, JavaScript, PHP 5 и других.

Смысл парсинга в том, чтобы ускорить рутинную работу. Чтобы собрать и сохранить в таблице список из тысячи статей на сайте, человек потратит часы. Парсер сделает эту работу за несколько минут. Парсер ускоряет работу в сотни раз и допускает меньше ошибок, чем человек.

**Что можно парсить и зачем это нужно**

Количество способов использования парсинга бесконечно. С помощью парсинга можно извлекать любую информацию из исходного кода страниц сайтов, из соцсетей, каталогов, файлов и документов.

Мы перечислим самые распространённые способы применения парсинга в маркетинге. Парсерами ищут:

* Цены на товары и услуги. Парсинг сайтов помогает быстро собрать информацию о ценах на продукты конкурентов. Если проводить его регулярно, можно всегда поддерживать конкурентную цену на продукт. Снижать её, когда конкуренты делают скидки, — и повышать, если их цены тоже растут.
* Поисковые фразы. Можно использовать парсинг выдачи «Яндекса», чтобы собрать все возможные ключевые слова. Парсер может сохранять фразы из поисковых подсказок и блока «Люди ищут» внизу.
* Целевую аудиторию в соцсетях. С помощью парсинга собирают людей по нужным критериям — например, тех, кто проявляет активность в сообществах конкурентов. Потом этим людям показывают рекламу. Сейчас очень популярен парсинг во «ВКонтакте», в других соцсетях его используют реже.
* Битые ссылки. Так называют ссылки, ведущие на несуществующую страницу. Парсер может быстро обойти весь сайт и сохранить в документе все битые ссылки — и страницы, на которых они расположены.
* Идеи для контента. Парсер может собрать в одном месте посты или статьи на выбранную тему. Это нужно, чтобы изучить сразу всех конкурентов и выпустить интересный контент. Некоторые парсят сайты, чтобы копировать контент, — например, добывают так описания для тысяч товаров в интернет-магазине.
* Отзывы. Если компания занимается [SERM](https://skillbox.ru/media/marketing/chto-marketologu-nuzhno-znat-o-serm/) — управлением репутацией в поисковых системах, она может использовать парсеры, чтобы находить новые отзывы. Тогда программы каждый день будут собирать отзывы, появившиеся за последние сутки, а менеджеры быстро отреагируют на них.