НПОУ «ЯКУТСКИЙ КОЛЛЕДЖ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ» (НПОУ «ЯКИТ»)

Отделение информационных технологий и туризма

Специальность 09.02.07 «Информационные системы и программирования»

по дисциплине:

ЯЗЫКИ ПРОГРАММИРОВАНИЯ

Веб-парсинг сайта Озон для сбора отзывов с использованием семантического анализа текста.

Исполнитель студент

гр. КИСП-23\_\_\_\_\_\_\_\_\_\_\_\_\_

                                                               подпись, дата

Руководитель,

преподаватель\_\_\_\_\_\_\_\_\_\_\_\_\_

                                                               подпись, дата

Якутск, 2024

Содержание

[ВВЕДЕНИЕ 3](#_Toc191723608)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 4](#_Toc191723609)

[1.1. Основы веб-парсинга 4](#_Toc191723610)

[1.2. Семантический анализ текста 5](#_Toc191723611)

[1.3. Анализ отзывов как источник информации 7](#_Toc191723612)

[ПРАКТИЧЕСКАЯ ЧАСТЬ 11](#_Toc191723613)

[ЗАКЛЮЧЕНИЕ 12](#_Toc191723614)

ВВЕДЕНИЕ

Ozon является одним из крупнейших маркетплейсов на территории СНГ, предлагающим широкий выбор товаров в различных категориях. Удобная система поиска и интуитивно понятный личный кабинет позволяют пользователям легко находить нужные товары и отслеживать свои покупки.

Семантический анализ текста — это процесс оценки содержательности текста и определения его значения. Этот метод анализа помогает понять, какие ключевые идеи и смыслы заложены в тексте.

Актуальность: Ozon - крупнейший маркетплейс на территории СНГ. Отзывы являются неотъемлемой частью любого маркетплейса. Они помогают, более детально изучить товар и подтолкнуть к его покупке или избежать некачественного товара.

Объект исследования: маркетплейс Ozon.

Метод исследования: веб-парсинг.

Цель: создание веб-парсера для сбора отзывов с платформы Ozon, чтобы проанализировать их содержание и выявить полезные инсайты для улучшения качества сервиса и повышения удовлетворенности покупателей.

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

* 1. Основы веб-парсинга

Веб-парсинг (или парсинг данных) — это процесс автоматизированного извлечения информации с веб-сайтов. Этот метод позволяет собирать данные, которые обычно представлены в виде текста, изображений и других форматов, преобразуя их в структурированный вид для дальнейшего анализа или использования.

Для выполнения веб-парсинга используются специальные программы или скрипты, которые имитируют действия пользователя, загружают страницы сайта, анализируют HTML-код и извлекают нужные данные. Веб-парсинг может применяться для различных целей, таких как мониторинг цен, сбор новостей, создание баз данных и многое другое.

Важно отметить, что веб-парсинг должен осуществляться с учетом законодательства и правил использования контента каждого конкретного сайта. Некоторые сайты могут запрещать автоматическое извлечение данных через свои условия использования или использовать технические меры защиты от скрейпинга.

Для начала необходимо получить доступ к содержимому веб-страницы. Это делается с помощью HTTP-запросов. Популярные библиотеки для этого:

Python:

* Requests: самая популярная библиотека для отправки HTTP-запросов.
* Urllib: встроенная библиотека для работы с URL.
* Httpx: Современная асинхронная библиотека для работы с HTTP-запросами.
* Aiohttp: Асинхронная библиотека для работы с HTTP-запросами, которая использует asyncio.

Каждая из этих библиотек имеет свои особенности и подходит для разных задач. Например, если вам нужно просто отправить запрос и получить ответ, то requests будет идеальным выбором. Если же вы работаете над высоконагруженным проектом, где важна скорость и возможность параллельного выполнения запросов, стоит рассмотреть использование httpx или aiohttp.

Методы парсинга:

* Ручной парсинг

Это самый простой метод, когда вы вручную копируете данные с сайта. Подходит только для небольших объемов информации и не подходит для автоматизации.

* Автоматический парсинг с использованием регулярных выражений

Регулярные выражения позволяют находить и извлекать текстовые шаблоны из HTML-кода страницы. Этот метод эффективен при работе с простыми структурами данных, но становится сложным при обработке сложных веб-страниц.

* Парсинг с помощью CSS-селекторов

Многие библиотеки и фреймворки поддерживают работу с CSS-селекторами, что позволяет легко выбирать нужные элементы на странице. Например, document.querySelector в JavaScript или функции селекции в библиотеках вроде BeautifulSoup.

* XPath

XPath – это язык запросов к XML-документам, который также используется для работы с HTML. С его помощью можно извлекать информацию из структурированных документов, включая сложные структуры страниц.

* 1.5. API

Если сайт предоставляет API, то использование этого метода будет наиболее предпочтительным, так как он обеспечивает доступ к данным напрямую без необходимости парсить HTML-код.

* 1. Роли

Проект был распределён среди участников команды следующим образом:

1. Разработкой кода для парсинга занимаются Левин Артем и Сафаргалеев Владимир.

План выполения работы для разработки кода:

* Фа
* Фы
* Фы

1. Разработкой Telegram бота Исааков Илья

План выполения работы для бота:

* Фа
* Фы
* Фы

1. Разработкой искусственного интеллекта Варламов Никита

План выполения работы для искусственного интеллекта:

* Фа
* Фы
* Фы

1. Разработкой базы данных Огонёрова Сардаана и Яковлева Евгения

План выполения работы для разработки базы данных:

* Фа
* Фы
* Фы
  1. Библиотеки (СТЕК технологий)

Для веб-парсинга на Python можно использовать следующий стек технологий:

Для получения данных. Библиотека Requests для выполнения HTTP-запросов к веб-ресурсам.

Для разбора данных. Библиотека BeautifulSoup для парсинга (разбора) данных в HTML и XML форматах.

Для обработки данных и базового анализа. Библиотека Pandas для обработки данных в условно табличном формате.

Также для более масштабного веб-парсинга в Python можно использовать фреймворк Scrapy. Он эффективно обрабатывает все этапы веб-парсинга, предоставляя класс spider для получения данных и конвейеры для их обработки.

Ещё одна полезная библиотека — Selenium. Это средство автоматизации веб-тестирования, которое можно использовать для целей парсинга. Оно позволяет взаимодействовать с веб-страницами на стороне браузера и делать запросы, нажимая на определённые элементы или кнопки

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

ЗАКЛЮЧЕНИЕ

Использование результатов анализа отзывов открывает перед бизнесом широкие возможности для улучшения всех аспектов своей деятельности – от разработки новых продуктов до оптимизации внутренних процессов. Главное – внимательно слушать своих клиентов и своевременно реагировать на их обратную связь.