|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **НПОУ «ЯКУТСКИЙ КОЛЛЕДЖ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ»** | | | | | | | |
| (наименование образовательной организации) | | | | | | | | | |
|  | | | | | | | | | |
|  | **ОТДЕЛЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ** | | | | | | |
| (наименование структурного подразделения (кафедра / отделение)) | | | | | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
| **ОТЧЕТ** | | | | | | | | | |
| **по Языкам программирования** | | | | | | | | | |
| Веб-парсинг сайта Озон для сбора отзывов с использованием семантического анализа текста | | | | | | | | | |
| (наименование типа практики) | | | | | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
|  | | |  | Выполнил: | | | | | |
|  | | |  | обучающиеся | | КИСП-23 | группы | | |
|  | | |  | Варламов Никита Андреевич  Исаков Илья Игоревич  Левин Артем Данилович  Огонерова Сардаана Васильевна  Сафаргалеев Владимир Викторович  Яковлева Евгения Николаевна | | | | | |
|  | | |  | (фамилия, имя, отчество (при наличии)) | | | | | |
|  | | |  | \_\_\_\_\_\_\_\_\_\_\_ «\_\_\_» \_\_\_\_\_\_\_\_ 20\_\_\_ г. | | | | | |
|  | | |  | (личная подпись, дата) | | | | | |
|  | | |  |  | | | | | |
|  | | |  |  | | | | | |
| Проверил руководитель практики | | |  | Проверил преподаватель: | | | | | |
|  | | |  |  | | | | | |
| (уч. степень, уч. звание, должность) | | |  | (уч. степень, уч. звание, должность) | | | | | |
|  | | |  | Федоров Дьулуур Андрианович | | | | | |
| (фамилия, имя, отчество (при наличии)) | | |  | (фамилия, имя, отчество (при наличии)) | | | | | |
| Оценка | |  |  | Оценка |  | | | | |
| \_\_\_\_\_\_\_\_\_\_\_ «\_\_\_» \_\_\_\_\_\_\_\_ 20\_\_\_ г. | | |  | \_\_\_\_\_\_\_\_\_\_\_ «\_\_\_» \_\_\_\_\_\_\_\_ 20\_\_\_ г. | | | | | |
| (личная подпись, дата) | | |  | (личная подпись, дата) | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
|  | | | | | | | | | |
| Якутск, 2025 | | | | | | | | | |

Содержание

[ВВЕДЕНИЕ 3](#_Toc195354893)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 4](#_Toc195354894)

[1.1. Терминология по проекту и глоссарий 4](#_Toc195354895)

[1.2. Распределение ролей и работы 6](#_Toc195354896)

[1.3. Стек технологий 7](#_Toc195354897)

[2. ПРАКТИЧЕСКАЯ ЧАСТЬ 8](#_Toc195354898)

[ЗАКЛЮЧЕНИЕ 9](#_Toc195354899)

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.

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.

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.

Объектом исследования процесс веб-парсинга сайта Озон для сбора отзывов с использованием семантического анализа текста, а методом исследования -  веб-парсинг и семантический анализ.

Целью проекта является создание веб-парсера для сбора отзывов с платформы Ozon для анализа тональности.

Задачи:

* Подготовка
* Парсинг страниц с отзывами
* Семантический анализ текста
* Визуализация результатов

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

* 1. Терминология по проекту и глоссарий

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.

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.

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.

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

Терминология:

* Семантика: наука о значениях знаков и символов.
* Значимость: конкретные смыслы отдельных элементов текста.
* Контекст: окружающая ситуация, влияющая на понимание значения.
* Антецедент: исходный элемент, от которого зависит другое слово или выражение.
* Консистенция: согласованность смыслов внутри текста.
* Энтитизация: выделение сущностей (объектов, лиц, мест и др.).
* Инференция: вывод новых фактов из имеющихся данных.
* Кореференция: установление связи между разными частями текста, относящимися к одному объекту.
* Омонимия: наличие разных значений одного слова.
* Гипонимия/гиперонимия: отношения типа "род-вид" ("яблоко" гипоним "фрукта").
* Синтаксичность: взаимосвязь синтаксиса и семантики.
  1. Распределение ролей и работы

Проект был распределён среди участников команды следующим образом:

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

План выполнения работы для разработки кода:

1. Изучение материала

* Использование библиотек(Panda)

1. Процесс написание кода

* Что как делали

1. Отладка
2. Разработкой Telegram бота Исаков Илья

План выполнения работы для бота:

* Изучение материала
* «Фы» Написание кода
* Отладка

1. Разработкой искусственного интеллекта Варламов Никита

План выполнения работы для искусственного интеллекта:

* Изучение материала
* Нахождение примеров
* Создание кода
* Отладка

1. Разработкой базы данных Огонёрова Сардаана и Яковлева Евгения

План выполнения работы для разработки базы данных:

* Изучение материала
* Выбор СУБД, создание базы данных
* Отладка
  1. 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.