ОБЗОР И РАЗРАБОТКА ПРОГРАММЫ ПО СБОРУ ЭЛЕКТРОННЫХ РЕСУРСОВ НА КАЗАХСКОМ ЯЗЫКЕ ПО ЗАКОНОДАТЕЛЬСТВУ РЕСПУБЛИКИ КАЗАХСТАН

Айтбеккызы Аружан
КазНУ имени аль-Фараби.

Шормакова Асем
КазНУ имени аль-Фараби.

Аннотация: Работа посвящена обзору программ по сбору электронных ресурсов. В условиях стремительной цифровизации средств обучения, культурных архивов и коммуникаций, вопрос эффективного сбора, хранения и использования языковых данных становится ключевым для сохранения культурного наследия и поддержания языкового разнообразия. В исследовании проведен литературный обзор, охватывающий актуальные аспекты электронных ресурсов, языковых данных и программ сбора данных. Целью работы является разработка программы, специфичной для казахского языка. В процессе исследования решаются задачи изучения законодательной базы, анализа существующих программ и разработки универсальной системы сбора данных на казахском языке.
Ключевые слова: электронные ресурсы, казахский язык, краулинг.

Annotation: The work is devoted to the development of a program for collecting electronic resources. With the rapid digitalization of teaching media, cultural archives and communications, the issue of effective collection, storage and use of language data becomes key to the preservation of cultural heritage and the maintenance of linguistic diversity. The study conducted a literature review covering relevant aspects of electronic resources, language data and data collection programs. The goal of the work is to develop a program specific to the Kazakh language. In the process of research, the tasks of studying the legislative framework, analyzing existing programs and developing a universal data collection system in the Kazakh language are solved.
Keywords: electronic resources, Kazakh language, crawling.

Введение

Растущий спрос на данные на казахском языке в сфере научных исследований, образования, культуры и бизнес-аналитики. Разработка не только удовлетворит потребности в доступе к информации на национальном языке, но и соответствует законодательству о защите данных и языковом многообразии. Законодательство Республики Казахстан отражает стремление к гармоничному сочетанию традиций и современности, а также поддержанию и развитию казахского языка в цифровой эпохе. Программа сбора электронных ресурсов призвана создать эффективный механизм для агрегации, хранения и доступа к разнообразным материалам на казахском языке, включая литературные произведения, исследования, языковые ресурсы и другие материалы. Таким образом, создание программы по сбору электронных ресурсов на казахском языке отражает стремление к сохранению, развитию и цифровой адаптации своего уникального языкового и культурного наследия в эпоху современных технологий.

Литературный обзор

В современном информационном обществе электронные ресурсы играют ключевую роль в сохранении и развитии языков и культур. В связи с этим, разработка программ по сбору электронных ресурсов казахского языка в соответствии с законодательством Республики Казахстан становится необходимой задачей. Целью данного литературного обзора является рассмотрение актуальных исследований и публикаций, касающихся этой темы, с целью обоснования необходимости разработки подобной программы. Статья "Аналитическая обработка текстовых ресурсов и документов на казахском языке" авторов Рахимовой и других (2019) посвящена разработке интеллектуальных поисковых систем, способных искать и извлекать новую информацию из текстовых данных на казахском языке в области образования. Авторы обосновывают актуальность данной темы в связи с растущим объемом данных в цифровой форме, предоставляющих возможность доступа к различным источникам электронных документов. Применение интеллектуальных поисковых систем позволит удовлетворить информационные потребности пользователей, и поэтому разработка информационно-аналитических поисковых систем, позволяющих работать с данными на казахском языке, является актуальной.

Цель исследования заключается в разработке эффективных алгоритмов и моделей для интеллектуальных поисковых систем на основе современных технологий в области информационного поиска и обработки естественного языка. Авторы стремятся создать системы, способные проводить анализ текстовых ресурсов на казахском языке, выявлять новые знания и улучшать качество образования в соответствии с потребностями современного информационного общества. Таким образом, данная работа имеет важное значение для развития интеллектуальных поисковых систем в контексте использования казахского языка и образования.

Работа Рахимова и Сатыбалдиева (2020) посвящена разработке системы автоматического сбора и обработки открытых данных на казахском языке из сети интернет, что имеет практическую значимость в задачах сбора и анализа текстов. Введение работы обосновывает актуальность выбранной темы, проводится обзор существующих подходов и формулируются задачи исследования. Одной из центральных задач является сбор и первичная обработка текстовых данных с последующим анализом. Авторы подчеркивают, что сбор данных является первоочередной задачей из-за неструктурированности открытых данных в сети интернет, требующей дополнительной обработки [1-2].

В работе представлена система обработки веб-страниц казахско-язычных порталов, а также приведено практическое применение данного подхода на реальных данных открытых ресурсов с помощью созданной системы. В рамках исследования представлен подход индексирования документов с использованием признаков. Ожидается, что разработанная система поможет структурировать открытые данные с ресурсов интернета на казахском языке и провести анализ собранных данных.

Методы сбора электронных ресурсов

1.Веб-краулинг (Web Crawling) [3]:

Веб-краулинг — это процесс автоматического сканирования веб-страниц для извлечения информации.

Веб-кроулинг, также известный как веб-сканирование или веб-спайдинг, — это автоматическое переползание (навигация) по веб-страницам в Интернете с целью сбора

информации. Краулеры или веб-пауки — это программные роботы, которые просматривают веб-страницы и собирают с них данные для различных целей, таких как индексация содержимого для поисковых систем, анализ структуры веб-страниц, извлечение информации и другие цели.

Процесс веб-ползания начинается с предоставления краулеру URL-адреса веб-страницы, которую необходимо посетить. Затем краулер загружает содержимое страницы и анализирует его, извлекая ссылки на другие страницы. Затем краулер переходит по этим ссылкам и продолжает процесс переползания, повторяя его до тех пор, пока не переполнит весь набор страниц или не будут выполнены другие условия остановки.

Веб-краулеры широко используются для различных целей, в том числе:

1. индексирование содержимого для поисковых систем, чтобы пользователи могли находить информацию в Интернете.

2. сбор данных для анализа, исследования или отслеживания онлайн-активности.

3. анализ структуры веб-сайтов для оптимизации производительности или улучшения качества работы пользователей.

4. проверка ссылок на веб-сайтах или обнаружение неработающих ссылок.

5. сканирование контента в целях безопасности или обнаружения угроз.

На языке программирования Python можно использовать библиотеки, такие как Scrapy или BeautifulSoup, для написания краулеров, которые могут автоматически переходить по ссылкам и извлекать данные с веб-страниц.

2.Использование API:

API (Application Programming Interface) - это интерфейс, позволяющий различным программным приложениям взаимодействовать друг с другом. API определяет методы и правила взаимодействия между различными программными компонентами, позволяя им обмениваться данными и выполнять определенные функции.

Основными особенностями API являются:

- Структурированные запросы и ответы: API определяет структуру запросов, которые могут быть отправлены программному обеспечению, и формат ответов, получаемых в ответ на эти запросы.
- Набор методов и функций: API предоставляет набор методов или функций, которые могут быть вызваны для выполнения определенных действий или операций. Эти методы и функции обычно документируются, чтобы разработчики могли понять их работу и использование.
- Стандартизация взаимодействия: использование API стандартизирует способ взаимодействия приложений друг с другом, облегчая интеграцию и обмен данными между различными системами.

Примерами API являются веб-интерфейсы, обеспечивающие доступ к данным или услугам через Интернет, API операционных систем, обеспечивающие управление ресурсами компьютера, API приложений, обеспечивающие доступ к функциям приложения, и т. д.

Многие онлайн-сервисы предоставляют API для доступа к своим данным и функциональности.

На языке Python можно использовать библиотеки для работы с API, например, requests для выполнения HTTP-запросов. Более подробную информацию можно увидеть в таблице 1.

Таблица1. Особенности методов сбора данных

Можно еще перечислить следующие методы сбора:

3.Веб-скрапинг [4]:

Веб- скрапинг — это процесс автоматического сбора информации с веб-страниц для извлечения данных. Веб-скрапинг выполняется с помощью программных средств, называемых веб-очистителями, которые получают доступ к веб-страницам, загружают их содержимое и извлекают нужные данные.

Основными компонентами веб-срапинг являются:

1. Загрузка веб-страницы: веб-скребки загружают HTML-код веб-страницы с помощью HTTP-запросов.

2. Анализ HTML: после загрузки страницы веб-скребки анализируют HTML-код страницы для поиска релевантных данных. Это может включать поиск определенных тегов (пример, или <p>) или атрибутов (пример, класс или идентификатор).

3. Извлечение данных: найденные HTML-элементы анализируются и извлекаются данные, соответствующие указанным критериям.

4. Обработка данных и резервное копирование: извлеченные данные могут быть обработаны и сохранены в нужном формате, например, в базе данных, CSV или JSON-файле.

Приложения для веб-анализа могут быть самыми разными - от автоматического сбора цен на товары в интернет-магазинах до сбора заголовков новостей с различных новостных сайтов. Однако важно отметить, что веб-анализ должен выполняться в соответствии с руководящими принципами использования Интернета и законами о защите данных. На некоторых веб-сайтах действуют правила, запрещающие несанкционированное удаление их содержимого, поэтому перед выполнением веб-очистки важно убедиться, что правила и политика этого веб-сайта не противоречат этим правилам [5-6].

Использование библиотек для анализа HTML-кода веб-страниц и извлечения нужных данных.

Метод сбора данных	Преимущества	Недостатки
Веб-краулинг	Автоматизация процесса	учет правовых и этических норм
Использование API	Доступ к данным через официальные источники	Ограниченный доступ
Веб-скрапинг	Гибкость и адаптивность	Нарушение правил использования данных
Скачивание и архивирование	Удобство хранения и передачи	Ограниченный доступ к определенным источникам

Используя и анализируя методы сбора анализа, программа включает в себя элементы веб-краулинга (для загрузки и отображения веб-страницы) и веб-скрапинга (для анализа HTML-кода и извлечения информации, такой как текстовые блоки из тегов). Более подробнее показана на рисунке 1.

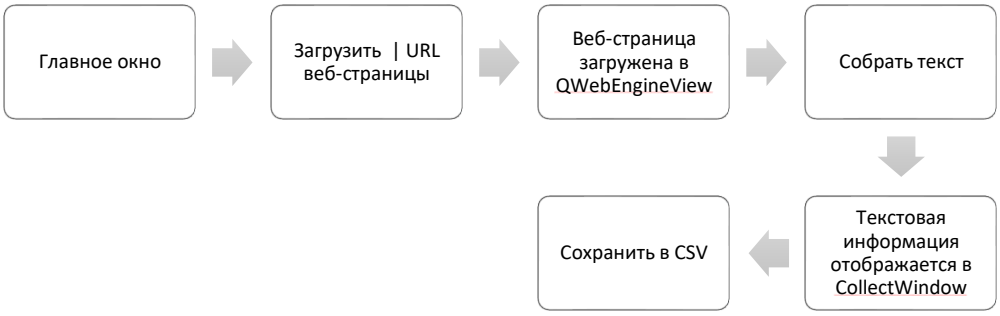

Рисунок 1. Общая схема программы по сбору данных.

Начало программы выглядит так (Рисунок 2)

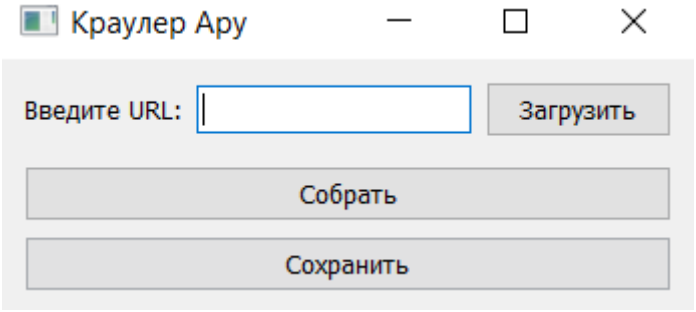

Рисунок 2. Начало работы программы

После нажатия кнопки “Собрать” программа собирает интернет-ресурсы с указанных источников.

Вывод

В заключении надо сказать, что автоматический сбор данных для любого языка очень необходим. Количество данных и объем нужен большой так как для обработки требуется большая статистика. Поэтому если владеть современными нынешними инструмента сбора данных, то можно облегчить задачу. И данная предложенная программа актуальна для юридических документов на казахском языке.

Использованная литература:

1. Shormakova, Zh. Zhumanov, B. Abduali, D. Rakhimova and D. Amirova Analytical Processing of Textual Resources and Documents in the Kazakh Language // Journal of Engineering and Applied Sciences 14 (20): 7714-7721, 2019
2. Д.Р. Рахимова, А.Р. Сатыбалдиев АЛГОРИТМ СБОРА ТЕКСТОВЫХ ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ // «Физика-математика ғылымдары» сериясы, №2(70), 2020
3. Okogwu, Flora Ifeoma, "Understanding Electronic Resources Collection Development Practices Through Selected Theories" (2021). // Library Philosophy and Practice (e-journal). 5704.
4. Краулер, интернет ресурс (ru-brightdata.com)
5. Владимир Дронов, Николай Прохоренок // Python 3 и PyQt 5. Разработка приложений (2016)

6. Abdurakhmonova, N., & Ismailov, A. S. (2022). APPLYING WEB CRAWLER TECHNOLOGIES FOR COMPILING PARALLEL CORPORA AS ONE STAGE OF NATURAL LANGUAGE PROCESSING. In СОВРЕМЕННАЯ ФИЛОЛОГИЯ. СОЦИАЛЬНАЯ И НАЦИОНАЛЬНАЯ ВАРИАТИВНОСТЬ ЯЗЫКА И ЛИТЕРАТУРЫ (pp. 22-27).
7. https://primeminister.kz/decisions