ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ

әл-Фараби атындағы Қазақ ұлттық университетінің

ХАБАРЛАРЫ

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ
НАУК РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный
университет имени аль-Фараби

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
al-Farabi Kazakh National University

SERIES
PHYSICO-MATHEMATICAL

3 (343)

JULY – SEPTEMBER 2022

PUBLISHED SINCE JANUARY 1963

PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *ақпараттық коммуникациялық технологиялар сериясы.*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Типографияның мекен-жайы: «Аруна» ЖК, Алматы қ., Мұратбаев көш., 75.

«Известия НАН РК. Серия физика-математическая».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75.

News of the National Academy of Sciences of the Republic of Kazakhstan.
Physical-mathematical series.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Address of printing house: ST «Aruna», 75, Muratbayev str, Almaty.

NEWS OF THE NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 3, Number 343 (2022), 5-18
https://doi.org/10.32014/2022.2518-1726.137
УДК 28.23.29
МРНТИ 28.17.27

А.С. Аканова[1*], А.А. Макашев[1], С.А. Наурызбаева[1], Н.Н.Оспанова[2]

[1]Казахский агротехнический университет им. С. Сейфуллина,
Казахстан, Астана;

[2]Павлодарский университет имени С. Торайгырова,
Казахстан, Павлодар.

E-mail: akerkegansaj@mail.ru

МОДЕЛИРОВАНИЕ ТЕМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ДАННЫХ ИЗ ИНТЕРНЕТА

Аннотация. Процесс подготовки для результативного извлечения данных из Интернета по различным тематикам сталкивается с проблемой структурирования и организации процесса поиска данных и и их извлечения. Для решения данной проблемы можно успешно применить моделирование действий, производимых во время поиска и извлечения информации из Интертнета. Были исследованы веб-парсинги с разных предметных областей, таких как финансовые данные, пси хологические исследования и другие. Описаны особенности работы веб-парсеров, способы хранения собранных данных. Иссле-до ваны понятия применяемые в области извлечения данных с Интернета. Также, в статье говорится о смежных темах, таких как NLP, глубокое и машинное обучение, и как они непосредственно связаны с процессом парсинга. В статье приведена модель поиска и извле че ния текста из Интернета, работа программы описывается в виде диаграммы прецедентов и диаграммы активности. Данные диаграммы используются в первоначальном выполнении проекта и описании требований заказчика аналитиком. Для упрощения работы разработчика приме-няются различные виды диаграмм, но в большинстве случаев удобно использовать выше названные диаграммы для моделирования прог-рам много продукта. Также приведена схемаработы метода doc2bow,

который используется в машинном обучении при извлечения текста по темам. Также проведен обзор на современные инструменты парсинга, работающие с языком программирования Python. А именно библиотека BeautifulSoup, фреймворк Scrapy и набор инструментов для автоматизации тестирования Selenium. В конечном результате, были построены UML-диаграммы модели, которые подробно показывают процесс веб-парсинга. Представленная модель извлечения данных из Интернета является визуализацией действий производимых приложением. Предлагаемая диаграмма может использоваться при разработке приложений по извлечению данных из Интернет ресурса.

Ключевые слова: веб-парсинг, веб-парсер, моделирование, извлечение данных, диаграмм активности.

А.С. Ақанова[1*], А.А. Макашев[1], С.А. Наурызбаева[1], Н.Н. Оспанова[2]

[1]С. Сейфуллин атындағы Қазақстан агротехникалық университеті, Қазақстан, Астана;
[2]С. Торайғыров атындағы университеті, Қазақстан, Павлодар.
E-mail: akerkegansaj@mail.ru

ИНТЕРНЕТТЕН ТАҚЫРЫП БОЙЫНША ДЕРЕКТЕРДІ АЛУДЫ МОДЕЛДЕУ

Аннотация. Интернеттен әртүрлі тақырыптар бойынша деректерді тиімді алуға дайындық процесі деректерді іздеу және оларды алу процесін құрылымдау және ұйымдастыру проблемасына тап болады. Бұл мәселені шешу үшін Интертнеттен ақпаратты іздеу және алу кезінде жасалған әрекеттерді модельдеу арқылы әр-түрлі жолдарды тиімді қолдануға болады. Қаржылық деректер, психологиялық зерттеулер және басқалар сияқты әртүрлі тақырыптардағы веб-парсингтер зерттелді. Веб-парсерлердің ерекшеліктері ол жиналған деректерді сақтау әдістері арқылы сипатталды. Интернеттен деректерді алу саласында қолданылатын ұғымдар зерттелді. Сондай-ақ, мақалада NLP, терең және машиналық оқыту сияқты байланысты тақырыптар және олардың талдау процесіне тікелей қатысы туралы айтылады. Мақалада интернеттен мәтінді іздеу және шығару моделі берілген, бағдарлама прецеденттер диаграммасы және белсенділік диаграммасы түрінде сипатталған. Бұл диаграммалар жобаның бастапқы орындалуында және тапсырыс берушінің талаптарын сипаттап әзірлеушіге дайындау

үшін аналитик (талдаушы) жұмыс жасайды. Әзірлеушінің жұмысын жеңілдету үшін әртүрлі диаграммалар қолданылады, бірақ көп жағдайда бағдарламалық өнімді модельдеу үшін жоғарыда аталған диаграммаларды қолдану ыңғайлы. Сондай-ақ, тақырып бойынша мәтін шығару кезінде машиналық оқытуда қолданылатын doc2bow әдісінің сызбалары келтірілген. Python бағдарламалау тілімен жұмыс істейтін заманауи талдау құралдарына шолу жасалды. Атап айтқанда, BeautifulSoup кітапханасы, Scrapy шеңбері және Selenium тестін автоматтандыруға арналған құралдар жиынтығы туралы ақпарат берілген. Нәтижесінде веб-талдау процесін егжей-тегжейлі көрсететін UML диаграммалары арқылы модель жасалды. Интернеттен деректерді шығарудың ұсынылған моделі қолданба жасаған әрекеттерді визуализациялау болып табылады. Ұсынылып отырған диаграмма Интернет ресурстан деректерді шығару бойынша қосымшаларды әзірлеу кезінде пайдаланылуы мүмкін.

Түйін сөздер: веб-парсинг, веб-парсер, модельдеу.

A.S. Akanova[1*], A.A. Makashev[1] , C.A. Наурызбаева[1], N.N. Ospanova[2]

[1]Kazakh agrotechnical university, Kazakhstan, Astana;
[2]S.Toraigyrov University, Kazakhstan, Pavlodar.
E-mail: akerkegansaj@mail.ru

MODELING OF THEMATIC DATA EXTRACTION FROM THE INTERNET

Abstract. The process of preparing to extract data from the Internet on various topics faces the problem of structuring and organizing the data retrieval process. To solve the problem, modeling of actions performed when searching and extracting information from the Internet is used. Web parsings from such subject areas as financial data, psychological research and others were investigated. The features of the work of web parsers, methods of storing the collected data are described. The concepts used in the field of data extraction are investigated. The article describes NLP, deep and machine learning and presents a model for searching and extracting text from the Internet. The work of the program is described in the form of a precedent diagram and an action diagram. Diagrams are used by the analyst at the beginning of the project to outline the customer's requirements. The model of the doc2bow method, which is used in machine learning when

extracting text by topic, is shown. There is also an overview of parsing tools, namely, the BeautifulSoup library, the Scrapy framework and a set of tools for automating Selenium testing. As a result, UML diagrams of the model were built, which show in detail the process of web parsing. A model of data extraction actions from the Internet is presented. The proposed model can be used in the development of applications for data extraction.

Key words: web-parsing, web-parser, parser, parsing, modeling.

Введение. В эпоху накопления и обработки больших данных большое место в науке имеют исследования процесса извлечения данных из открытых Интернет источников. Такие технологии как NLP (Natural Language Processing) – обработка естественного языка требуют огромного количества данных, для того чтобы эффективно применить алгоритмы машинного обучения. В связи с требованием большого количества данных, не только для NLP, но и других сфер, появились инструменты для извлечения данных, которые называются «веб-парсеры», «парсеры», либо «веб-скрейперы». Инструменты извлечения данных, занимают особое место и в сфере психологических исследований (Speckmann, 2021). Каждое действие человека в Интернете оставляет «следы»: посты, комментарии, понравившиеся статьи и т.д. Speckman F. отметил, что такие «следы» непременно помогут в психологических исследованиях, поскольку каждый «след» описывает поведение человека. Инструментов извлечения текстов кроме применения в психологических исследованиях, использовались и в сфере финансов (Krotov et al., 2018). В статье показан сбор данных с помощью написания на языке программирования R веб-парсера. Термин «парсинг» обозначает синтаксический анализ. Веб-парсер можно запустить с помощью написания специального скрипта, например, на языке Python. При написании скрипта, задаются необходимые условия, для извлечения определенных блоков текста. Тем самым, посредством некоторых предустановленных правил производится анализ текста, в случае с веб-сайтами это весь HTML-документ. Далее, текст, который попал под условие скрипта необходимо сохранить. Для хранения можно использовать как обычные текстовые файлы, так и базы данных (Mahmood et al., 2018), о том, как делать сбор данных в какое-либо хранилище данных. Таким образом, с помощью веб-парсера можно автоматизировать сборку огромного количества данных, экономя на этом ресурсы и время. На сегодняшний день актуальность и достоверность информации являются приоритетом, что своей работой и

выполняет парсер. Путем запуска скрипта по расписанию, можно всегда получать свежую информацию.

Одной из особенностей при написании парсеров является то, что имеется возможность собирать данные не с одного веб-сайта, а с нескольких. Гибкость работы парсера напрямую зависит от разработчика, который пишет скрипт.

Извлечение данных, это систематизированный процесс извлечения и комбинирование необходимой информации из веб-ресурсов (Glez-Peña et al, 2014). Насколько нам известно веб-парсер имитирует действия человека при извлечении данных с веб-сайтов. А также информацию можно извлекать только с открытых веб-ресурсов.

В последнее время становится актуальным применение технологий глубокого обучения при анализе текста (Rekha et al., 2022). В статье сравниваются традиционные методы извлечения текста от сложных процессов. При извлечении текста из Интернета важно применение синтаксического анализа, в том числе универсальной зависимости, который рассматривается с помощью метода извлечения текста с межязыковыми отношениями (Taghizadeh et al., 2022). Синтез и извлечение структурированных данных с помощью неглубокого синтаксического анализа и сегментации предложений были предложены учеными и имел успех при работе с рускоязычным патентом (Korobkin et al., 2019). Веб-парсинг страниц можно отнести к более традиционным методам, поскольку процесс парсинга обходится без использования сложных технологий таких как нейронные сети и глубокое обучение. Но эффективное использование технологий веб-парсинга также не является простым. Для того, чтобы добиться нужного результата необходимо разработать корректную модель, отстроить архитектуру приложения, что в итоге будет вести себя согласно ожидаемым результатам.

Одним из примеров извлечения текста с источников является научная работа (Frisoni et al., 2021). В работе описывается то, как извлечение текста из публикаций, поможет справиться с их постоянным ростом. Необходимость извлечения полезной, структурированной информации является одной из основ веб-парсинга. Для этого необходимо определить четкие правила в алгоритме работы скрипта, применять регулярные выражения и корректные условия. Анализ HTML-разметки, а также использование XPath выражений являются актуальными и надежными при использования регулярных выражений (Antonov et al., 2020). Регулярные выражения являются неким шаблоном с определенными заданными внутри условиями. С помощью этих выражений можно найти нужные строки или подстроки. XPath запросы часто используются при

веб-парсинге. С помощью них можно находить нужные в DOM-структуре элементы, с которых далее будет происходить вычитка данных.

На данный момент не существует прямого законодательного органа, который бы занимался мониторингом веб-парсеров (Silva et al., 2018), но существуют нормативные документы в котором защищены авторские права. Существует еще одно правило «этики» при запуске веб-парсера на веб-сайт. Веб-парсер не должен нагружать сервер, на котором находится веб-сайт, то есть необходимо соблюдать интервал, при котором происходит парсинг, чтобы не навредить серверу веб-сайта.

Самой популярной библиотекой на языке Python является BeautifulSoup (Vargiu et al., 2013). С помощью данной библиотеки, можно написать парсер статичных страниц. Библиотека не имеет возможности парсить данные с динамически подгружающихся страниц, что является главным недостатком данной библиотеки.

В статье (Sirisuriya et al., 2015) говорится о том, что большинство веб-сайтов не дают доступа сохранить копию данных. Такая проблема действительно существует, поскольку многие веб-сайты используют для рендера страниц JavaScript или AJAX, который в свою очередь подгружает весь контент страницы динамический с помощью асинхронных запросов к серверу. Существуют библиотеки, которые решают данную проблему. Одна из них написана на языке Python: scrapy-splash. Это фреймворк, которые дает возможность создавать парсеры способные собирать данные, даже если веб-сайт используют загрузку контента через JavaScript или AJAX.

Многие крупные веб-сайты, такие как Twitter, iTunes, TikTok и др. предоставляют специальные API интерфейсы для того, чтобы получать определенные данные с их веб-сайта. В Twitter это посты, в iTunes информацию о треке, его продолжительность, наименование, альбом и т.д., в TikTok получение информации о профиле пользователя, искать пользователей и посты. Но у этих API интерфейсов существуют различные ограничения: требуется авторизация, ограничение на кол-во запросов в определенный интервал времени. О том как парсить данные веб-сайты в обход этим ограничениям описано в статье (Hernandez-Suarez., 2018). Он предложил использовать вышеописанную библиотеку написанную для Python – Scrapy. Движок scrapy отправляет запрос по ссылке поиска постов, также передаются необходимые параметры. Движок получает ответ от сервера и отправляет полученную страницу HTML элементами загрузчику. Далее, паук (в фреймворке Scrapy их принято так называть) начинает анализировать и парсить необходимые данные, которые в конце попадают в определенное хранилище данных.

Веб-парсинг используется для собра текстовой информации, которая в свою очередь требует обработки. Одной из проблем в компьютерной лингвистике является извлечение из текста информации по каким-либо темам.

Следовательно, исследование в области веб-парсинга требуют разработки модели извлечения информации, для добычи нужных данных, то есть избежать как можно меньше «мусора» при извлечении данных.

Для разработки модели извлечения информации необходимо обозначить следующие задачи:

- выполнить обзор и анализ научных статей
- разработать модель сбора целевых данных при веб-парсере на основе языка UML

Материалы и методы исследования. Моделирование извлечения данных по тематике. В данном исследование были применены такие методы исследования как анализ, синтез, сравнение.

Для того, чтобы понять, как работает парсер, и откуда собираются данные необходимо ознакомиться с понятием DOM-дерево (Document Object Model). DOM-дерево полностью показывает структуру веб-документа и иерархию всех HTML-элементов. Каждый элемент DOM-дерева – это то, что видит пользователь, находясь на веб-сайте: кнопка, меню, баннеры, картинки, видео (Uzun, 2020). Все эти элементы на странице считаются объектами. Эти элементы называют «тегами». По иерархии теги делятся на два вида – родительские и дочерние (Донова и др., 2019).

Основными действиями при извлечении данных по тематике является моделирование приложения. Для моделирования процесса веб-парсинга необходимо использовать UML (Unified Modeling Language) – унифицированный язык моделирования.

Чтобы определить какие роли непосредственно участвуют в процессе веб-парсинга и каким поведением они обладают, нужно смоделировать Use-case диаграмму (диаграмма прецедентов). Модель, которой показана на рисунке 1.

Рисунок 1. Use-case диаграмма (диаграмма прецедентов)

В диаграмме прицедентов (рисунок 1) показано распределение основных ролей объектов при извлечении данных. Основными исполнителями являются веб-сайт, сервер на котором работает веб-сайт, и, собственно, сам веб-парсер (инструмент, с помощью которого извлекают данные). Веб-парсер получает нужную страницу и начинает извлекать из нее данные которые требуются. Сервер отвечает за хранение БД. Веб-сайт отображает данные, приходящие с сервера. Между ними есть действие, которое связывает их – получение и отправка запросов. С помощью запросов веб-сайт и веб-парсер получают данные с сервера, в котором все данные и хранятся. После успешного выполнения запроса на получение содержимого HTML-страницы, происходит парсинг страницы нужных блоков данных.

Для того, чтобы подробно понять логику поведения системы необходимо построить диаграмму активностей (Activity diagram), которая показана на рисунке 2.

В диаграмме активностей можно подробно рассмотреть алгоритм действий веб-парсинга. Здесь можно наблюдать каждый шаг, перед тем как придет ожидаемый результат в виде нужного текста. Данная диаграмма показывает какая роль у объектов (веб-парсер, веб-сайт и сервер.) и за что они ответственны в системе.

Рассмотрев, вышеописанные UML-модели, разработчику не сложно определиться этапами разработки программного продукта предназначенный для тематического извлечения данных.

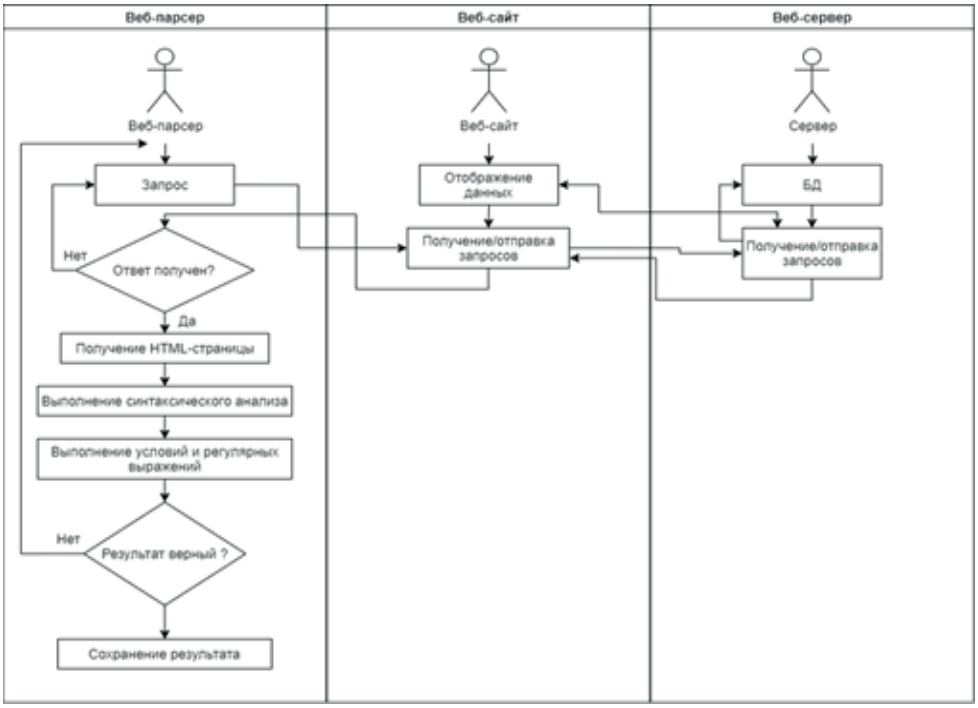

Рисунок 2. Activity диаграмма (диаграмма активностей)

Данная модель помогает решить некоторые проблемы разработчика:

1) Визуально увидеть полноценную картину программного продукта для аврсинга

2) Определиться с классами, которые ему потребуются в работе

3) Определить связи между объектами и их действия.

4) Легкая модернизация программного продукта

5) Выявление ошибко при тестирование

6) Выделение основных задач перед разработкой программного продукта

Архитектура программного продукта состит из веб-парсера (это может быть инструмент, либо приложение) и хранилища данных. Веб-парсер делает запрос к веб-сайту, получает ответ в виде HTML-страницы, далее производится синтаксический анализ по условиям, которые указал пользователь, в поиске нужных элементов помогают BeautifulSoup, Selenium или Scrapy, с помощью регулярных выражений можно повысить точность поиска определенных слов, фраз в тексте. Далее полученные результаты сохранять в БД, таблицы или текстовые файлы, для дальнейшей работы с ними.

При работе с веб-парсингом нужно помнить, что парсинг страниц на веб-сайтах, где имеется конфиденциальная информация, обход запретов на копирование контента с веб-сайтов будет являться нарушением закона. Необходимо соблюдать «этику» веб-парсинга, не нагружать сервера множеством параллельных запросов.

Элементы можно искать с помощью тега, класса, либо id элемента, также по CSS-селектору. Класс тега – это идентификатор, который используется при задании каких-либо стилей с помощью CSS. С помощью поиска через класс тега, можно сузить область поиска необходимого нам элемента, тем самым ускоряя веб-парсинг. Id – это уникальный идентификатор, который используется для наложения стилей на тег с помощью CSS, либо при наложении JavaScript скрипта на тег. CSS-селектор – специальный синтаксис, который используется в css файлах, при указывании тегов. Например: body p – означает, что внутри тега «body», необходимо найти тег «p»; p.paragraph-1 – означает, что необходимо найти тег p с классом «paragraph-1». С помощью определения данных условий парсинг находит нужные для сбора данных элементы.

Результаты. Из исследованных статьей следует отметить, что авторы использовали две библиотеки веб-парсинга – BeautifulSoup и Scrapy. Представляем описание некоторых характеристик библиотек веб-парсинга (таблица 1) и приемлемую библиотеку для применения его в программном продукте.

Таблица 1 - Характеристика BeautifulSoup, Scrapy и Selenium

Критерий\ Наименование парсера	BeautifulSoup	Scrapy	Selenium
Размер библиотеки	Легковесная библиотека, которая не требует много свободного места	Поскольку Scrapy является фреймворком для веб-парсинга, то занимает больше места	Selenium изначально был разработан для автоматизации тестирования веб-приложений, поэтому требует много места
Производительность	Данная библиотека работает медленно, но с применением технологии многопоточности производительность повышается	Scrapy является очень быстрым парсером, поскольку оптимизирован для извлечения больших объемов данных	Selenium является быстрым инструментом для парсинга, но уступает Scrapy

Парсинг JavaScript и AJAX страниц	BeautifulSoup не имеет возможности парсить страницы подгруженные динамически с помощью JavaScript или AJAX	У Scrapy есть возможность парсить динамически подгруженные страницы путем добавления библиотеки scrapy-splash	Selenium умеет парсить страницы подгружаемые JavaScript-ом, поскольку это приложение для автоматизации тестирования веб-приложений такая функция в нем имеется
Гибкость	BeautifulSoup имеет множество зависимостей, из-за чего его сложнее масштабировать	Scrapy легко масштабируется, имеет множество функционала для того чтобы увеличивать поток приходящих данных	Selenium может работать на больших проектах, но требует тщательной настройки ограничения лимита данных
Кроссплатформенность	Имеется кроссплатформенность поскольку это библиотека Python	Имеется кроссплатформенность поскольку это библиотека Python	Имеется кроссплатформенность
Документация	BeautifulSoup является популярным среди сообщества и у него хорошо расписанная, доступная для новичков документация	Scrapy имеет более сложную и менее детальную документацию, что является сложностью для новичков	Selenium имеет хорошо расписанную документацию, которая будет понятна как новичкам, так и опытным разработчикам
Доступность	Данная библиотека является рекомендацией для начинающих разработчиков веб-парсеров	Данная библиотека требует опыта от разработчика, используется в больших и сложных проектах	Данный инструмент требует опыта от разработчика

Из вышеописанной таблицы 1, можно сделать вывод, что BeautifulSoup и Scrapy написаны на языке Python, а значит он свободно распространяется. Selenium использует несколько инструментов. BeautifulSoup является библиотекой основанной на множестве зависимостей, что делает ее менее гибкой, не позволяет легко масштабироваться, но имеет упрощенный порог вхождения для разработчиков. Scrapy – фреймворк, разработанный для полноценного выполнения крупных процессов веб-парсинга и используется в больших проектах. Порог вхождения высокий. Selenium является «средним» между BeautifulSoup и Scrapy. С помощью него можно быстро парсить веб-страницы основанные на JavaScript.

Моделирование, применяя абстракцию при представлении данных показывает информацию в доступ визуальном формате. Тематическое извлечении текста на языке моделирования может выглядеть в виде сигналов, знаков, картинок и других объектов визуализации. Извлекая текст по ключевым словам из Интернет, можно определить их тематику и распределить их по «мешкам» или иначе применить метод Doc2bow из библиотеки Gensim. Doc2bow обычно используется в обучении данных. Слова объединяются в одну тематику (или как говорят оказываются в одном мешке) с помощью выбора слов с частотой образования друг с другом биграмм или триграмм.

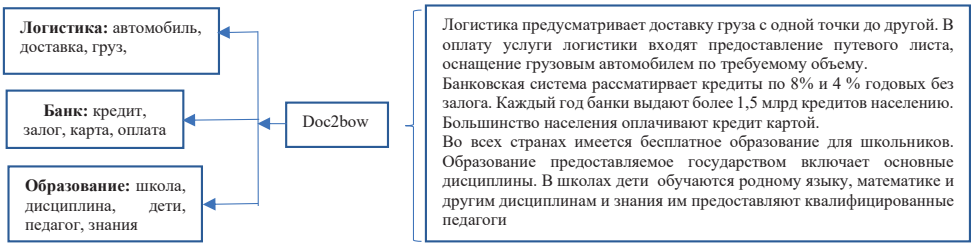

Рисунок 3. Модель схема работы Doc2bow

Отсюда, для тематического извлечения данных с Интернет ресурса предлагаем UML-диаграммы модели (рисунок 1, рисунок 2), которые подробно показывают процесс веб-парсинга.

Обсуждение. Представленные в статье диаграмма прецедентов и диаграмма активити являются одними из самых популярных видов диаграм, используемых в моделирование бизнес процессов программных продуктов. В данном случае данная гепотеза коррелирует с результатами исследований Донова М.М., Лозина Е.Н., Веретенникова Е.Г. и Барклаевской Н.В. (Донова и др., 2019, Барклаевская, 2015). Учитывая интенсивное развитие IT в Республике Казахстан была предложена одна из перспективных направлений проектной работы – моделирование, в котором решаются проблемы и совершенствования взаимодействия объектов программного продукта. Чем, обоснована решающая роль моделирования в разработке програмных продуктов. Перспективами дальнейших исследований являются применение концептуальной модели проекта. Предложен оптимальный инструмент для извлечения данных из Интернет ресурсов.

Итого, в результате были выполнены обзор и анализ научных статей, была разработана модель извлечения данных при веб-парсере на основе языка UML.

Заключение. Основными проблемами моделирования в разработке

программного продукта является отсутсвие аналитика либо аналитического мышления и не знание диаграмм языка UML. В результате данной работы была построена диаграмма прецедентов, в котором выделены актеры и их действие, что является ключевым моментом для разработчика. Разработчик в свою очередь может выделить себе без затруднений классы и методы в них. Диаграмме активности показывает последовательность работы выполнения действий актерами (классами).

Таки образом, моделирование извлечение текста из Интернета представлена в виде двух диаграмм на языке UML и схемы-модели работы doc2bow, что позволяет не только разработчику наглядно видеть процесс, но и заказчику.

Information about authors:

Akanova Akerke Saparovna – S. Seifullin Kazakh agrotechnical university, PhD, Senior lecturer, +77054480680, akerkegansaj@mail.ru; ORCID ID: http://orcid.org/0000-0003-2783-186X;

Makashev Ahmadi – S. Seifullin Kazakh agrotechnical university, Master student, +77086749153, *ahmadi98ahmadi@gmail.com*;

Nauryzbaeva Saya Amanzholovna – S. Seifullin Kazakh agrotechnical university, Master of Informatic, assistant lecturer, + 77016831139, ORCID ID: https://orcid.org/my-orcid?orcid=0000-0002-8544-0528;

Ospanova Nazira Nurgazyevna – Toraighyrov University, Associate Professor, +77011664573, nazira_n@mail.ru; ORCID ID: http://orcid.org/0000-0001-9416-0993.

REFERENCES:

Biegert J., Hofer H., Schultz A.H. (2015) A comparative study on web scraping. 86 DOI:10.1159/000433586 (in Eng).

Antonov E., Lopatina E., Ionkina K., Tretyakov E. Agent data merging Procedia Computer Science Том 169, Страницы 473 - 4782020 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 Seattle15 August 2019 DOI 10.1016/j. (procs.2020.02.).

Frisoni G., Moro G., Carbonaro A. A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave IEEE Access Открытый доступ Том 9, Страницы 160721 – 1607572021 DOI 10.1109/ACCESS.2021.3130956 (in Eng).

Glez-Peña D., Lourenço A., López-Fernández H., Reboiro-Jato M., Riverola F. F. (2014) Web scraping technologies in an API world (2014) // Briefings in bioinformatics, 15(5) 788-797. https://www.semanticscholar.org/paper/Web-scraping-technologies-in-an-API-world-Glez-Peña-Lourenço/b4951eb36bb0a408b02fadc12c0a1d8e680b589f (in Eng).

Hernandez-Suarez A. et al. Hernandez-Suarez Aldo, Sánchez-Pérez G., Toscano-

Medina K., Martínez-Hernández V., Sanchez Victor, Meana H. (2018) A web scraping methodology for bypassing twitter API restrictions. *https://arxiv.org/pdf/1803.09875.pd* (in Eng).

Korobkin D.M., Vasiliev S.S., Fomenkov S.A., Lobeyko V.I. (2019) Extraction of structural elements of inventions from Russian-language patents. 4th International Conference on Big Data Analytics, Data Mining and Computational Intelligence and the 8th International Conference on Theory and Practice in Modern Computing, Porto, DOI: 10.33965/tpmc2019_201907l020 (in Eng). Krotov V., Tennyson M. (2018) Research note: scraping financial data from the web using the R language. Journal of Emerging Technologies in Accounting, 15 (1), 169-181. DOI: https://doi.org/10.2308/1558-7940-15.1.i (Published: 01 July 2018) (in Eng).

Krotov V., Silva L. (2018) Legality and ethics of web scraping. Americas Conference on Information Systems 2018: Digital Disruption. https://www.researchgate.net/publication/324907302 Legality_and_Ethics_of_Web_Scraping (in Eng).

Mahmood A., Khan H.U., Alarfaj F.K., Ramzan M., Ilyas M. (2018) A multilingual datasets repository of the Hadith content. International Journal of Advanced Computer Science and Applications, 9 (2), 165 – 172. DOI 10.14569/IJACSA.2018.090224 (in Eng).

Rekha D., Sangeetha J., Ramaswamy V. (2022) Digital document analytics using logistic regressive and deep transition-based dependency parsing. Journal of Supercomputing, 78 (2), 2580 – 2596, DOI 10.1007/s11227-021-03973-4 (in Eng).

Speckmann F. (2021). Web scraping: A useful tool to broaden and extend psychological research // Zeitschrift für Psychologie. 229(4), 241-247, DOI: 10.1027/2151-2604/a000470 (in Eng).

Taghizadeh N., Faili H. (2022) Cross-lingual transfer learning for relation extraction using Universal Dependencies Computer Speech and Language, 71, 101265. DOI 10.1016/j.csl.2021.101265 (in Eng).

Uzun E. (2020) A novel web scraping approach using the additional information obtained from web pages //IEEE Access. 8. 61726-61740. (in Eng).

Vargiu E., Urru M. (2013) Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artifical Intellegent Resours, 2 (1), 44-54. (in Eng).

Barclayevskaya N.V. The use of the unified modeling language UML in the project approach when teaching students in the specialty “business informatics. Materials of the scientific and methodological conference of the SZIU RANEPA, RANEPA under the President of the Russian Federation, Moscow. *https://elibrary.ru/download/elibrary_25659228_48815743.pdf.*

Dontsova M.M., Lozina E.N., Veretennikova E.G. Analysis and modeling of tutors’ work processes. Problems of design, application and security of information systems in the digital economy Materials of the XIX International Scientific and Practical Conference. Rostov-on-Don. https://elibrary.ru/download/elibrary_43785016_98237740.pdf.

DOM tree. [electronic resource]. URL: https://learn.javascript.ru/dom-nodes.

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein
the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Заместитель директор отдела издания научных журналов НАН РК *Р. Жәлиқызы*
Редакторы: *М.С. Ахметова, Д.С. Аленов*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 15.09.2022.
Формат 60х881/8. Бумага офсетная. Печать – ризограф.
17,5 п.л. Тираж 300. Заказ 3.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19