ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
INFORMATION TECHNOLOGY AND TELECOMMUNICATIONS

УДК 004.056
DOI:10.21822/2073-6185-2023-50-1-114-122 Оригинальная статья/ /Original Paper

Разработка программно-технического решения для выявления трендов спроса на товары

А.И. Мифтахова[1], Э.И. Янгиров[1], Е.И. Карасева[1], А.И. Янгиров[2], Е.Ю. Никулина[3], И.Г.Дровникова[3]
[1]Университет ИТМО,
[1]197101, г. Санкт-Петербург, Кронверкский проспект, д.49, лит. А, Россия,
[2]ФКУ «НИЦ «Охрана» Росгвардии,
[2]111539, г. Москва, Реутовская, 12Б, Россия,
[3]Воронежский институт МВД России,
[3]394065, г. Воронеж, пр. Патриотов, 53, Россия

Резюме. Цель. Целью исследовательской работы является разработка программного решения для выявления трендов спроса на потребительские товары путем анализа больших данных. **Методы.** Для достижения поставленной цели в работе было проанализировано текущее состояние развития рынка интернет-ретейла в России, а также рассмотрены технологии и инструменты анализа больших данных, необходимые для проектирования программно-технического решения. Для оценки эффективности полученной модели обработки данных применяется выборка, полученная из открытых источников. **Результат.** В результате исследования разработано техническое решение, позволяющее анализировать спрос на товары в заданном временном диапазоне на основе данных из открытых источников. **Вывод.** Разработан программный компонент, позволяющий анализировать спрос на потребительские товары на основе данных о заказах. Полученное техническое решение поддерживает пакетную обработку данных, а архитектура инфраструктурного компонента позволяет вести вычисления распределенно. Тестирование инструмента на реальной выборке показало эффективность такого подхода к анализу трендов потребительского спроса.

Ключевые слова: большие данные, спрос, тренд, естественный язык, программный компонент

Для цитирования: А.И. Мифтахова, Э.И. Янгиров, Е.И. Карасева, А.И. Янгиров, Е.Ю. Никулина, И.Г.Дровникова Разработка программно-технического решения для выявления трендов спроса на товары. Вестник Дагестанского государственного технического университета. Технические науки. 2023; 50(1):114-122. DOI:10.21822/2073-6185-2023-50-1-114-122

Development of a software and hardware solution to identify trends in demand for goods

A.I.Miftakhova[1], E.I.Yangirov[1], E.I.Karaseva[1], A.I.Yangirov[2], E.Yu.Nikulina[3], I.G. Drovnikova[3],
[1]ITMO University,
[1]49, lit. A, Kronverksky Ave., St. Petersburg 197101, Russia,
[2]FKU "Research Center "Protection" of the Russian Guard,
[2]12B Reutovskaya Str., Moscow 111539, Russia,
[3]Voronezh Institute of the Ministry of Internal Affairs of Russia,
[3]53 Patriotov Ave., Voronezh 394065, Russia

Abstract. Objective. The aim of the research work is to develop a software solution for identifying trends in demand for consumer goods by analyzing big data. **Method**. To achieve this goal, the work analyzed the current state of development of the Internet retail market in Rus-

sia, as well as the technologies and tools for analyzing big data necessary for designing a software and hardware solution. To evaluate the effectiveness of the obtained data processing model, a sample obtained from open sources is used. **Result.** As a result of the study, a technical solution has been developed that allows analyzing the demand for goods in a given time range based on data from open sources. **Conclusion.** A software component has been developed to analyze the demand for consumer goods based on order data. The resulting technical solution supports batch processing of data, and the architecture of the infrastructure component allows distributed computing. Testing the tool on a real sample showed the effectiveness of this approach to analyzing consumer demand trends.

Keywords: big data, demand, trend, natural language, software component

For citation: A.I. Miftakhova, E.I. Yangirov, E.I. Karaseva, A.I. Yangirov, E.Yu. Nikulina, I.G. Drovnikova. Development of a software and hardware solution to identify trends in demand for goods. Herald of the Daghestan State Technical University. Technical Science. 2023; 50(1):114-122. DOI:10.21822/2073-6185-2023-50-1-114-122

Введение. Выявление трендов спроса на товары играет ключевую роль в анализе продаж, помогает планировать закупки, актуализировать ассортимент, корректировать алгоритмы ценообразования и т.п. Существует множество способов выявления трендов, и один из них предполагает работу с большими массивами информации. Обычно они представляют собой потоки неструктурированных или слабоструктурированных данных и поступают в количествах, обработка которых невозможна обычными средствами.

Использование больших данных позволяет маркетологам компании получать достоверную информацию о текущем состоянии и тенденциях развития бизнеса, изучать поведение основных конкурентов, выявлять предпочтения своих клиентов.

Все это позволяет компании достигнуть конкретных результатов [1]: увеличение продаж; выявление наиболее популярных товаров и услуг; повышение качества обслуживания клиентов; уменьшение расходов и повышение рентабельности бизнеса; предупреждение мошенничества; удержание клиентов.

Постановка задачи. Целью исследования является разработка программного решения для выявления трендов спроса на потребительские товары путем анализа больших данных.

Методы исследования. В данном исследовании применяются следующие методы:

1) Анализ состояния рынка онлайн-торговли в России;

2) Выбор архитектуры для построения систем обработки больших данных;

3) Разработка программной части технического решения поставленной проблемы;

4) Проверка работоспособности полученного решения на основе больших данных (датасета), содержащего тестовые данные для анализа.

Обсуждение результатов. Анализ состояния рынка интернет-ретейла в России. Ситуация на рынке интернет-ретейла в России за последние годы претерпела сильные изменения. Пандемия COVID-19 спровоцировала бурный рост продаж, совершаемых в сети.

Объем рынка интернет-торговли в России за 2021 г. вырос на 52% и составил 4,1 трлн. руб., что следует из отчета аналитической компании Data Insight [2]. Число заказов выросло более чем вдвое - до 1,7 млрд., средний чек одной покупки снизился на 26% до 2400 руб., что представлено на рис. 1.

Снижение размера среднего чека связано с превращением онлайн-заказов в повседневную практику, опережающим ростом универсальных маркетплейсов с низким средним чеком, а также с ростом продаж в сегменте продуктов питания и смещением потребительского поведения в сторону небольших импульсных покупок с быстрой доставкой. Статистика данного показателя по годам представлена на рис. 2.

В целом доля электронной торговли на российском рынке ритейла за прошлый год увеличилась на 3 пп. и составила 12%.

Рис.1. Динамика количества заказов в онлайн-магазинах России с 2011 по 2021 гг.
Fig. 1. Dynamics of the number of orders in online stores in Russia from 2011 to 2021

Таким образом, рынок интернет-ретейла в России переживает бурный рост, что, несомненно, вызовет высокий уровень конкуренции в данной среде.

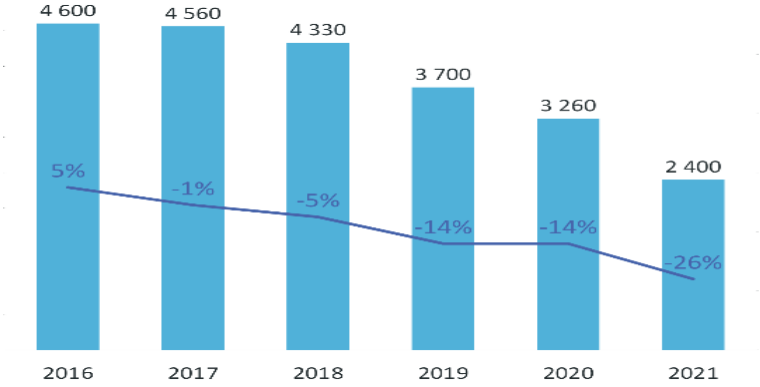

Рис.2. Динамика среднего чека покупок в онлайн-магазинах с 2016 по 2021 гг.
Fig. 2. Dynamics of the average check of purchases in online stores from 2016 to 2021

Следовательно, рост конкуренции на рынке интернет-товаров, в свою очередь обусловливает необходимость качественного анализа ситуации на данном рынке.

Разработка программно-технического решения для выявления трендов спроса на товары. Указанный выше анализ обосновывает необходимость создания специального инструмента выявления трендов спроса на потребительские товары. В рамках данного исследования были проведены разработка и тестирование программно-технического решения для анализа больших данных о потребительском спросе на рынке интернет-ретейла.

Одним из классических подходов к построению систем обработки больших данных (далее Big Data) является так называемая лямбда-архитектура [3]. Это концепция, согласно которой обработка информации происходит на двух основных уровнях:

1. Пакетный, на котором все входящие данные хранятся в необработанном виде, а затем обрабатываются в пакетном режиме. Обычно пакетный уровень представлен озерами больших данных (Data Lake) на базе Apache Hadoop [4]. В рамках данной статьи будет рассмотрена реализация именно этого способа обработки данных.

2. Потоковый, отвечающий за обработку информации в режиме реального времени. Этот уровень обеспечивает минимальную задержку передачи данных, но за счет снижения точности обработки данных [5].

Кроме того, выделяют также сервисный уровень (или уровень обслуживания), который индексирует пакеты и обрабатывает результаты вычислений, происходящих на пакетном уровне. Потоковый уровень обновляет уровень обслуживания, отправляя ему добавочную информацию с учетом последних данных. Преимуществами лямбда-архитектуры являются отказоустойчивость и масштабируемость [6]. Благодаря функциям пакетного и потокового уровней можно добавлять новые данные в основное хранилище, обеспечивая при этом сохранность существующих данных. В связи с этим лямбда-

архитектура широко используется на практике во многих Big Data проектах, в частности, в Twitter, Netflix и Yahoo [7]. Устройство лямбда-архитектуры представлено на рис. 3.

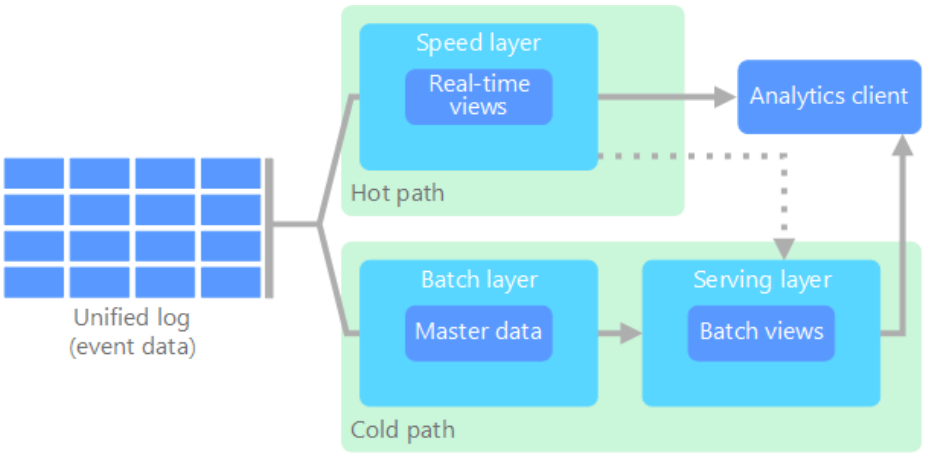

Рис. 3. Устройство лямбда-архитектуры
Fig. 3. Lambda architecture device

Реализация пакетного уровня обработки данных требует построения эффективной системы распределенного хранения и параллельных вычислений. В ходе изучения существующих решений для анализа больших данных такая архитектура зарекомендовала себя как наиболее отказоустойчивая и удобная для горизонтального масштабирования [8].

Данный подход предполагает использование следующих инструментов:

1. Apache Spark. Мастер Apache Spark обрабатывает входящие запросы от клиента (Jupyter Notebook), распределяет рабочие нагрузки и отслеживает ресурсы в нодах. Рабочие узлы выполняют работу, для которой они были выделены, и возвращают результат мастеру.

2. Для целей хранения данных предполагается интеграция с Apache Hadoop. Мастер в кластере Hadoop – это диспетчер ресурсов, который распределяет рабочую нагрузку по узлам кластера. NodeManager получает инструкции от ResourceManager и распределяет ресурсы, доступные на одном узле. Узлы данных отвечают за обслуживание запросов на чтение и запись.

3. Docker – программное обеспечение для автоматизации развёртывания и управления приложениями в средах с поддержкой контейнеризации, которое позволяет «упаковать» приложение со всем его окружением и зависимостями в контейнер [9].

4. В качестве клиентской части планируется использование Jupyter Notebook. Это среда разработки, располагающая широким функционалом для решения задач анализа данных и машинного обучения. Она позволяет разбить код на фрагменты и работать над ними в произвольном порядке – писать и проверять функции, загружать файлы в память и обрабатывать содержимое.

В качестве основного инструмента разработки программно-технического решения был выбран языка программирования Python. Он является наиболее популярным инструментом для анализа данных и машинного обучения [10]. Кроме того, данный язык обладает простым синтаксисом и позволяет разработчику сконцентрироваться на решении задачи, не вынуждая вдаваться в детали реализации.

Разработка указанного выше решения включает в себя создание так называемого пайплайна – документа, визуализирующего процесс разработки программно-технического продукта. Пайплайн обработки данных состоит из следующих этапов:

1. Стемминг – процесс сокращения слова до своей грамматической основы [11]. В качестве инструмента для проведения стемминга использовался модуль Snowball Stemmer библиотеки NLTK [12].

2. Удаление стоп-слов из всех исследуемых документов (так называемого корпуса документа) удаляются знаки препинания, числа, а также слова, не несущие смысловой нагрузки (предлоги, местоимения, союзы, междометия и пр.)

3. Токенизация – разбиение каждого документа в корпусе на токены-термы. В качестве токенизатора был использован модуль RegexpTokenizer из библиотеки NLTK [13].

4. Преобразование корпуса документов в матрицу TF-IDF. Это статистическая мера, используемая для оценки значимости терма в рамках документа, где TF – это отношение числа вхождений заданного слова в документ к общему количеству слов в нем, а IDF – это инверсия частоты встречаемости слова в рамках всего корпуса текстов. Учет IDF позволяет снизить вес слов, которые являются широко используемыми в рамках исследуемого массива данных. Таким образом, в рамках этого метода больший вес получают слова с высокой частотой употреблений в рамках одного документа и низкой – в рамках всего корпуса.

5. Кластеризация – процесс разбиения массива данных, представленного в векторном формате, на совокупность относительно однородных подмножеств [14]. Для реализации этого этапа был выбран алгоритм кластеризации, известный как метод k-средних. В частности, использована его реализация, представленная модулем KMeans в библиотеке scikit-learn [15].

6. Результаты кластеризации сохраняются в файл для обеспечения дальнейшего переиспользования без необходимости проводить повторные вычисления.

7. К размеченному на кластеры массиву применяется метод главных компонент (principal component analysis) из модуля RCA библиотеки Scikit-learn. Данный алгоритм призван уменьшить размерность входных данных, потеряв при этом наименьшее количество информации [16]. В рамках этой задачи производится сжатие массива векторов длины N до размерности 2. Этот этап необходим для визуализации и интерпретации результатов работы.

8. Для анализа смысловой нагрузки кластеров для каждого подмножества подбирается 10 слов, обладающих наиболее высоким значением TF-IDF. Таким образом, каждый кластер получает словесное описание, соответствующее общему смыслу входящих в него документов [17].

9. Полученные результаты выводятся на экран в виде точечной диаграммы. Для визуализации используется библиотека seaborn [18].

Обобщенную структуру пайплайна можно представить в виде, представленном на рис. 4.

Рис. 4. Пайплайн обработки входных данных
Fig.4. Input data processing pipeline

Тестирование программного компонента решения. В качестве датасета были использованы данные о заказах в онлайн-магазине одежды бренда H&M за период с 2019 – 2020 гг. [19]. Он содержит дату покупки и текстовое описание товара. Оригинальная вы-

борка содержит примерно 32 млн. записей. Массив данных записан в файл формата «*.parquet», обеспечивающий более эффективное с точки зрения занимаемого пространства хранение данных, чем стандартный формат «*.csv» [20].

Ниже представлен пример тестовых данных (рис. 5). В качестве первой тестовой выборки был выбран период с 01.12.2019 г. по 15.12.2019 г. Результаты анализа представлены на рис. 6.

	detail_desc	t_dat
1	Headband in a soft knit containing some wool with a draped detail at the front. Width approx. 1	2019-12-01
2	Fitted, padded jacket in woven fabric with a zip down the front and lined hood with a detachabl	2019-12-01
3	Boxy-style jumper in a soft, fine knit containing some wool with low dropped shoulders, long sl	2019-12-01
4	Short-sleeved cotton jersey top in a relaxed fit with a slightly wider neckline and a rounded her	2019-12-01
5	Belt in grained imitation leather with a large, patterned metal buckle. Width approx. 2.5 cm.	2019-12-01
6	Pyjama bottoms in lightweight fabric with an elasticated drawstring waist and tapered legs with	2019-12-01
7	Jumper in a soft, fine knit containing some wool with long sleeves, decorative metal buttons or	2019-12-01
8	Jacket in soft pile with a collar and zip down the front. Long raglan sleeves, pockets in the side	2019-12-01
9	Hoop earrings in metal decorated with pearly plastic beads. Length 6 cm.	2019-12-01
10	Set with a bra and thong briefs in satin with lace details. Soft, non-wired bra with triangular cup	2019-12-01
11	Sleeveless top in a viscose weave with a V-neck, adjustable spaghetti straps and embroidery a	2019-12-01
12	Joggers in soft jersey with wide elastication at the waist, side pockets and tapered legs.	2019-12-01
13	V-neck jumper in fine-knit cotton with ribbing around the neckline, cuffs and hem.	2019-12-01
14	Ankle-length trousers in twill made from a cotton blend. High paper bag waist with pleats and a	2019-12-01
15	Trousers in jersey with a high waist, concealed zip in one side and straight, wide legs with sew	2019-12-01
16	Short, thin metal chain necklace with leaf-shaped pendants. Adjustable length, 40-48 cm.	2019-12-01
17	Earrings with a round embossed metal disc and pearly pendants. Length 4.5 cm.	2019-12-01

Рис. 5. Фрагмент тестовых данных
Fig. 5. Test data snippet

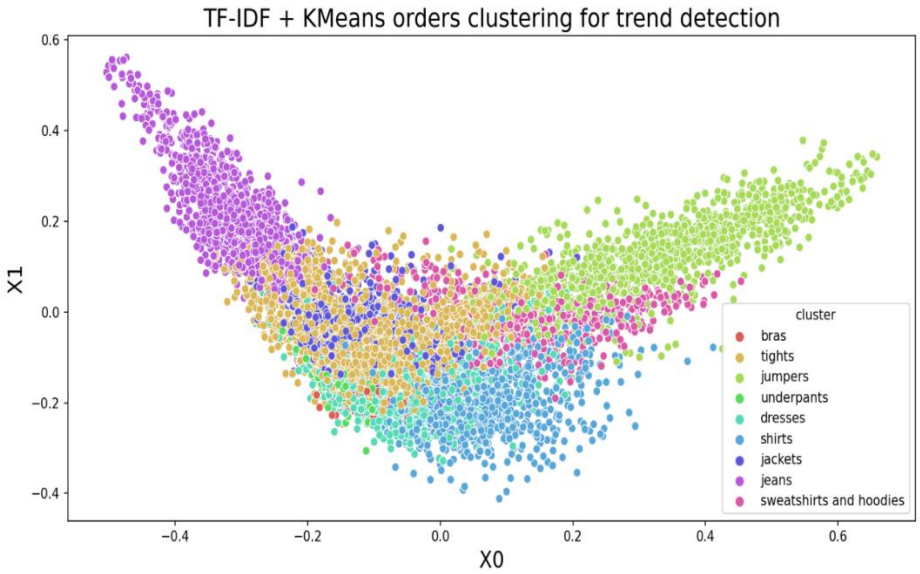

Рис. 6. Визуализация результатов обработки данных о заказах с 01.12.2019 г. по 15.12.2019 г.
Fig. 6. Visualization of the results of order data processing from 12/01/2019 to 12/15/2019

На графике, представленном на рис. 6, наиболее ярко выделяются кластеры заказов, содержащих свитеры, джинсы и колготки, что соответствует ожиданиям от покупок в зимний период. Можно отметить, что предметы одежды, выполняющие аналогичную функцию, сгруппированы по соседним кластерам (свитеры и худи).

Данный факт связан с тем, что описания данных товаров состоят из схожих наборов слов. Однако этим же можно объяснить и, на первый взгляд, необычные паттерны. Например, кластеры заказов, содержащих колготки и куртки, на графике практически смешиваются. Дело в том, что оба этих предмета одежды предназначены для утепления и могут быть описаны соответствующим набором слов (“warm”, “soft”, “thick”, “tight” и пр.). Для сравнения был рассмотрен летний период продолжительностью в 15 дней с 01.07.2019 г. по 15.07.2019 г. (рис. 7). Как видно из рис. 7, состав кластеров изменялся (появились, купальники и исчез кластер курток).

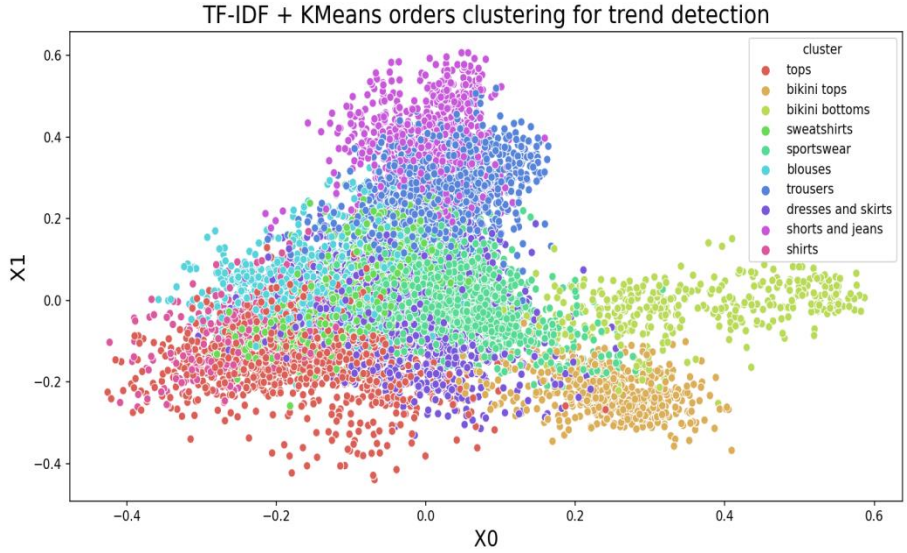

Рис. 7. Визуализация результатов обработки данных о заказах с 01.07.2019 г. по 15.07.2019 г.
Fig. 7. Visualization of the results of order data processing from 07/01/2019 to 07/15/2019

Очевидный интерес у покупателей вызывают пляжная одежда, шорты, джинсы, а также спортивная одежда, что также является ожидаемым потребительским поведением для летнего сезона.

Вывод. В рамках данного исследования был разработано программно-техническое решение, позволяющее анализировать спрос на потребительские товары на основе данных о заказах. Предложенное решение основано на пакетной обработке данных, а архитектура инфраструктурного компонента позволяет вести вычисления распределенно. Тестирование разработанного решения на реальной выборке показало эффективность такого подхода к анализу трендов потребительского спроса.

В качестве возможных задач для дальнейшей работы над данной темой могут быть выделены следующие:

1) Изучение инструментов для повышения точности и детальности определения трендов спроса на товары;

2) Реализация потокового модуля предложенной в исследовании лямбда-архитектуры;

3) Прогнозирование спроса на товары в заданной перспективе.

Библиографический список:

1. Величко, Н. А. Технология Big Data. Анализ рынка Big Data / Н. А. Величко, И. П. Митрейкин // Синергия Наук. – 2018. – № 30. – С. 937-943.
2. Черненко, О. С. Применение TF-IDF алгоритма в рекомендательных системах государственных закупок / О. С. Черненко // Мир компьютерных технологий: Сборник статей студенческой научно-технической конференции, Севастополь, 04–07 апреля 2017 года / Научный редактор Е.Н. Мащенко. – Севастополь: Федеральное государственное автономное образовательное учреждение высшего образования "Севастопольский государственный университет", 2017. – С. 66-67.
3. Леонтьева, С. А. Кластеризация изображений методом "k-средних" / С. А. Леонтьева, А. Ю. Демин // Молодежь и современные информационные технологии: Сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 03–07 декабря 2018 года/Томский политехнический университет. – Томск: Национальный исследовательский Томский политехнический университет, 2019. – С. 86-87.
4. Прохоренков, П. А. Современные информационные технологии маркетинга / П. А. Прохоренков, О. М. Гусарова, Т. В. Аверьянова // Фундаментальные исследования. – 2018. – № 12-1. – С. 158-162.
5. Маркетинговое исследование Интернет-торговля в России 2021 // Data Insight URL: https://datainsight.ru/eCommerce_2021 (дата обращения: 19.09.2022).
6. Kiran M. et al. Lambda architecture for cost-effective batch and speed big data processing //2015 IEEE International Conference on Big Data (Big Data). – IEEE, 2015. – С. 2785-2792.
7. Panwar A., Bhatnagar V. Data lake architecture: a new repository for data engineer //International Journal of Organizational and Collective Intelligence (IJOCI). – 2020. – Т. 10. – №. 1. – С. 63-75.

8. Григорьев Ю. А., Ермаков О. Ю. Обработка запросов в системе с лямбда-архитектурой на уровне ускорения //Информатика и системы управления. – 2020. – №. 2. – С. 3-16.
9. Матвеева П.Р. Сравнение лямбда и традиционной архитектуры //Форум молодых ученых. – 2018. – №. 1. – С. 734-740.
10. Fernández-Manzano E. P., Neira E., Clares-Gavilán J. Data management in audiovisual business: Netflix as a case study //El profesional de la información (EPI). – 2016. – Т. 25. – №. 4. – С. 568-576.
11. Big Data Solution with Hadoop, Spark, Jupyter and Docker // Medium URL: https://medium.com/@martinkarlsson.io/big-data-solution-with-hadoopspark-jupyter-and-docker-6763983ed5d8 (дата обращения: 24.09.2022).
12. Козинцев Д. А., Шиян А. А. Контейнеризация для анализа больших данных на примере kubernetes и docker //Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2020). – 2020. – С. 393-396.
13. Raschka S., Patterson J., Nolet C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence //Information. – 2020. – Т. 11. – №. 4. – С. 193.
14. Khyani D. et al. An Interpretation of Lemmatization and Stemming in Natural Language Processing //Journal of University of Shanghai for Science and Technology. – 2021.
15. URL: https://www.nltk.org/_modules/nltk/stem/snowball.html (дата обращения: 01.10.2022). Source code for nltk.stem.snowball//NLTK::nltk.stem.snowball
16. URL: https://www.nltk.org/_modules/nltk/tokenize/regexp.html (дата обращения: 01.10.2022). Source code for nltk.tokenize.regexp//NLTK::nltk.tokenize.regexp sklearn.cluster.KMeans // scikit-learn 1.1.2 documentation
17. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (дата обращения: 03.10.2022).
18. Granato D. et al. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective //Trends in Food Science & Technology. – 2018. – Т. 72. – С. 83-90.
19. Text Clustering with TF-IDF in Python // Medium URL: https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-pythonc94cd26a31e7 (дата обращения: 29.09.2022).
20. Seaborn: statistical data visualization // Seaborn Documentation URL: https://seaborn.pydata.org/index.html (дата обращения: 02.10.2022).
21. H&M Personalized Fashion Recommendations//Kaggle URL: https://www.kaggle.com/competitions/h-and-m-personalized-fashionrecommendations (дата обращения: 01.10.2022).
22. Saavedra M. Z. N., Yu W. E. A comparison between text, parquet, and PCAP formats for use in distributed network flow analysis on Hadoop //Journal of Advances in Computer Networks. – 2018. – Т. 5. – №. 2. – С. 59-64.

References:

1. Velichko N. A. Big Data technology. Analysis of the Big Data market/ N. A. Velichko, I. P. Mitreikin. *Synergy of Sciences*. 2018; 30: 937-943.[In Russ]
2. Chernenko O. S. Application of the TF-IDF algorithm in recommendatory public procurement systems-World of Computer Technologies: Collection of articles of the student scientific and technical conference, Sevastopol, April 04–07, 2017 / Scientific editor E .N. Mashchenko. - Sevastopol: Federal State Autonomous Educational Institution of Higher Education "Sevastopol State University", 2017; 66-67 [In Russ].
3. Leontieva, S. A. Clustering of images by the "k-means" method / S. A. Leontieva, A. Yu. Demin . Youth and modern information technologies: Proceedings of the XVI International scientific and practical conference of students, graduate students and young scientists , Tomsk, December 03–07, 2018 / Tomsk Polytechnic University. *Tomsk: National Research Tomsk Polytechnic University*, 2019; 86-87.[In Russ]
4. Prokhorenkov P. A. Modern information marketing technologies / P. A. Prokhorenkov, O. M. Gusarova, T. V. Averyanova. *Fundamental research.* 2018;12(1): 158-162.[In Russ]
5. Marketing research Internet commerce in Russia 2021. Data Insight URL: https://datainsight.ru/eCommerce_2021 (Accessed 09/19/2022). [In Russ]
6. Kiran M. et al. Lambda architecture for cost-effective batch and speed big data processing //2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015; 2785-2792.
7. Panwar A., Bhatnagar V. Data lake architecture: a new repository for data engineer. *International Journal of Organizational and Collective Intelligence (IJOCI).* 2020; 10(1): 63-75.
8. Grigoriev Yu. A., Ermakov O. Yu. Processing of requests in a system with lambda architecture at the level of acceleration. *Computer Science and Control Systems.* 2020; 2:3-16.[In Russ]
9. Matveeva P.R. Comparison of lambda and traditional architecture.*Forum of young scientists.* 2018;1: 734-740 [In Russ]
10. Fernández-Manzano E. P., Neira E., Clares-Gavilán J. Data management in audiovisual business: Netflix as a case study. *El profesional de la información (EPI)*. 2016;25(4): 568-576

11. Big Data Solution with Hadoop, Spark, Jupyter and Docker. Medium URL: https://medium.com/@martinkarlsson.io/big-data-solution-with-hadoopspark-jupyter-and-docker-6763983ed5d8 (Accessed: 09/24/2022)
12. Kozintsev D. A., Shiyan A. A. containerization for big data analysis on the example of kubernetes and docker. *Actual problems of infotelecommunications in science and education (APINO 2020).* 2020; 393-396. [In Russ]
13. Raschka S., Patterson J., Nolet C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence.*Information.* 2020;11(4):193.
14. Khyani D. et al. An Interpretation of Lemmatization and Stemming in Natural Language Processin. *Journal of University of Shanghai for Science and Technology*. 2021
15. URL: https://www.nltk.org/_modules/nltk/stem/snowball.html (Accessed: 01.10.2022). Source code for nltk.stem.snowball // NLTK:: nltk.stem.snowball
16. URL: https://www.nltk.org/_modules/nltk/tokenize/regexp.html (Accessed: 01.10.2022). Source code for nltk.tokenize.regexp // NLTK:: nltk.tokenize.regexp
17. URL:https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html(Accessed: 03.10.2022). sklearn.cluster.KMeans // scikit-learn 1.1.2 documentation
18. Granato D. et al. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology*. 2018;72:83-90.
19. Text Clustering with TF-IDF in Python // Medium URL: https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-pythonc94cd26a31e7 (Accessed: 29.09.2022).
20. Seaborn: statistical data visualization //Seaborn Documentation URL: https://seaborn.pydata.org/index.html (Accessed: 02.10.2022).
21. H&M Personalized Fashion Recommendations. Kaggle URL: https://www.kaggle.com/competitions/h-and-m-personalized-fashionrecommendations (Accessed: 01.10.2022).
22. Saavedra M. Z. N., Yu W. E. A comparison between text, parquet, and PCAP formats for use in distributed network flow analysis on Hadoop. *Journal of Advances in Computer Networks*. 2018; 5(2): 59-64.

Сведения об авторах:

Мифтахова Альбина Ирековна, студент, факультет инфокоммуникационных технологий; miftakhovaalbina@gmail.com

Янгиров Эмиль Илдарович, студент, факультет инфокоммуникационных технологий; emilyangirov@gmail.com

Карасева Екатерина Ивановна, кандидат экономических наук, доцент, факультета инфокоммуникационных технологий; eikaraseva@itmo.ru,

Янгиров Адиль Илдарович, начальник сектора функциональных испытаний инженерно-технических средств защиты отдела технических экспертиз и функциональных испытаний; adil-yan@yandex.ru

Никулина Екатерина Юрьевна, кандидат технических наук, доцент, кафедра автоматизированных информационных систем ОВД; 5nikeu@mail.ru

Дровникова Ирина Григорьевна, доктор технических наук, доцент, профессор кафедры автоматизированных информационных систем органов внутренних дел; idrovnikova@mail.ru

Information about the authors:

Albina I. Miftahova, Student; miftakhovaalbina@gmail.com

Emil I. Yangirov, Student; emilyangirov@gmail.com

Ekaterina I. Karaseva, Cand. Sci. (Econom), Assoc. Prof.; eikaraseva@itmo.ru

Adil I. Yangirov, Head of the sector of functional testing of engineering and technical means of protection of the department of technical expertise and functional tests; adil-yan@yandex.ru

Ekaterina Yu. Nikulina, Cand. Sci. (Eng), Assoc. Prof., Department of Automated Information Systems of the Department of Internal Affairs; nikeu@mail.ru

Irina G. Drovnikova, Dr. Sci. (Eng.), Prof., Assoc. Prof., Department of Automated Information Systems of Internal Affairs Bodies; idrovnikova@mail.ru

Конфликт интересов/Conflict of interest.

Авторы заявляют об отсутствии конфликта интересов/The authors declare no conflict of interest.

Поступила в редакцию/ Received 23.12.2022.

Одобрена после рецензирования / Reviced 24.01.2023.

Принята в печать /Accepted for publication 24.01.2023.