УДК 004.8+004.62

РЕСУРСЫ, ПРЕДОСТАВЛЯЮЩИЕ ДАННЫЕ ДЛЯ МАШИННОГО ОБУЧЕНИЯ И ПРОВЕРКИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Сикулер Денис Валерьевич

к.т.н., доцент кафедры «Информационные системы»,
e-mail: rimol@km.ru,
Российский государственный педагогический университет им. А. И. Герцена,
191186, г. Санкт-Петербург, набережная реки Мойки, д.48.

Аннотация. В статье выполнен обзор 10 ресурсов сети Интернет, позволяющих подобрать данные для разнообразных задач, связанных с машинным обучением и искусственным интеллектом. Рассмотрены как широко известные сайты (например, Kaggle, Registry of Open Data on AWS), так и менее популярные или узкоспециализированные ресурсы (к примеру, The Big Bad NLP Database, Common Crawl). Все ресурсы предоставляют бесплатный доступ к данным, в большинстве случаев для этого даже не требуется регистрация. Для каждого ресурса указаны характеристики и особенности, касающиеся поиска и получения наборов данных. В работе представлены следующие сайты: Kaggle, Google Research, Microsoft Research Open Data, Registry of Open Data on AWS, Harvard Dataverse Repository, Zenodo, Портал открытых данных Российской Федерации, World Bank, The Big Bad NLP Database, Common Crawl.

Ключевые слова: поиск данных, набор данных, открытые данные, репозиторий данных, каталог наборов данных, искусственный интеллект, машинное обучение.

Цитирование: Сикулер Д. В. Ресурсы, предоставляющие данные для машинного обучения и проверки технологий искусственного интеллекта // Информационные и математические технологии в науке и управлении. 2021. № 2 (22). С. 39-52. DOI:10.38028/ESI.2021.22.2.004

Введение. Разнообразные технологии и инструменты на основе искусственного интеллекта получают всё большее распространение и становятся более значимыми не только в различных сферах науки, техники, экономики и производства, но и в повседневной жизни обычных людей. Глобальная пандемия коронавируса COVID-19 продемонстрировала уязвимость человечества не только перед эпидемическими угрозами и сопутствующими проблемами, но и в аспекте вызванных ими долгосрочных последствий и ограничений [1-3]. Вместе с тем, пандемия еще раз показала и подтвердила, что успешно, эффективно и оперативно справляться с возникающими сложностями в различных областях деятельности можно, привлекая компьютерные средства и информационные технологии, в том числе связанные со сферой искусственного интеллекта [1, 3-12]. Многие проблемы и задачи, имеющие прямое или опосредованное отношение к эпидемии, будь то синтез вакцин и лекарств, анализ рентгеновских снимков, распределение больных и ресурсов и т.п., можно быстро и качественно решать лишь с применением интеллектуальных компьютерных технологий и инструментов [13-15]. Одним из ключевых факторов для разработки подобных эффективных технологий и средств является машинное обучение [7, 16-22]. Однако для успешного его осуществления, как правило, требуются многочисленные и разнообразные данные, которые во многих случаях может оказаться затруднительно собрать и подготовить [16, 17, 23-26]. В связи с этим особую роль приобретает поиск готовых данных. В [27] был рассмотрен ряд сайтов сети Интернет, предоставляющих доступ к различным данным, которые могут быть использованы с целью разработки и тестирования технологий и средств искусственного интеллекта и машинного обучения. Ниже приводится краткий обзор еще

нескольких ресурсов, с помощью которых можно найти данные для различных областей машинного обучения.

1. Kaggle (https://www.kaggle.com). На известном ресурсе, посвященном различным аспектам научно-исследовательской обработки данных [19, 21-23, 26, 28, 29], можно найти несколько десятков тысяч наборов данных, касающихся самых разнообразных сфер деятельности и прикладных задач (рис. 1). Например, представлено множество данных, относящихся к таким областям, как экономика и бизнес, классификация и распознавание образов, медицина и здоровье, образование и сфера развлечений, обработка изображений и естественного языка. Отдельно стоит отметить, что на сайте имеется более 2000 наборов данных, так или иначе касающихся COVID-19. С набором данных могут быть связаны так называемые блокноты (в том числе, Jupyter Notebooks), в которых можно выполнить интересующую обработку, а также исследовательские задачи, для которых можно предлагать решения в виде блокнотов. Таким образом, помимо собственно данных Kaggle предоставляет полноценную среду для работы с ними, включая инструменты на основе Jupyter. В блокнотах можно использовать программный код на языках Python или R. Поддерживаются различные форматы файлов данных, в том числе CSV, JSON и SQLite.

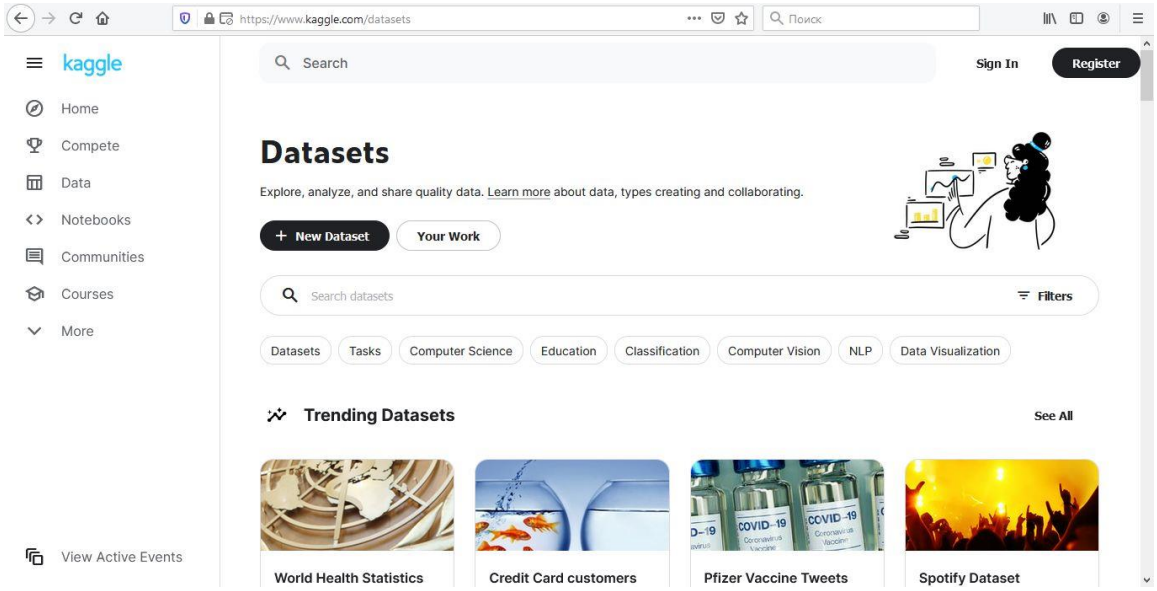

Рис. 1. Каталог наборов данных на ресурсе Kaggle

Для полноценной работы с ресурсом, в том числе для добавления и скачивания данных, необходима регистрация. Однако без регистрации можно просматривать данные и связанные с ними блокноты, а также создавать новые блокноты и выполнять с помощью них эксперименты (время работы ограничено).

На Kaggle также доступны так называемые состязания (конкурсы), в рамках которых нужно предложить решение для определенной задачи, связанной с обработкой данных [21, 22]. Для некоторых конкурсов предлагается финансовое вознаграждение за то или иное итоговое место.

Kaggle также предоставляет программный интерфейс (API) и инструменты командной строки для взаимодействия с ресурсом. Например, можно скачивать или создавать наборы данных, а также создавать и запускать на выполнение блокноты.

2. Google Research (https://research.google/tools/datasets/). На сайте исследовательского подразделения Google доступен каталог (рис. 2), включающий несколько десятков ресурсов с данными, которые можно использовать для различных изысканий, так или иначе связанных с компьютерными науками. В основном представлены данные, относящиеся к обработке:

- аудио; например, снабженная транскрипцией записанная речь на различных языках, в том числе малораспространенных типа баскского (Basque multi-speaker speech) или галисийского (Galician multi-speaker speech);
- текста; например, различные данные, извлеченные из страниц Wikipedia, в том числе Dictionaries for linking Text, Entities, and Ideas - база данных понятий (несколько миллионов) и связанных с ними слов и ссылок (более 100 миллионов);
- видео; например, YouTube-BoundingBoxes – сотни тысяч фрагментов видео из YouTube, которые снабжены миллионами аннотаций выделенных объектов различных классов (более 20) на кадрах;
- изображений; например, Quick Draw - многомиллионная коллекция ручных черно-белых рисунков-набросков, изображающих разнообразные предметы и понятия, относящиеся более чем к 300 категорий.

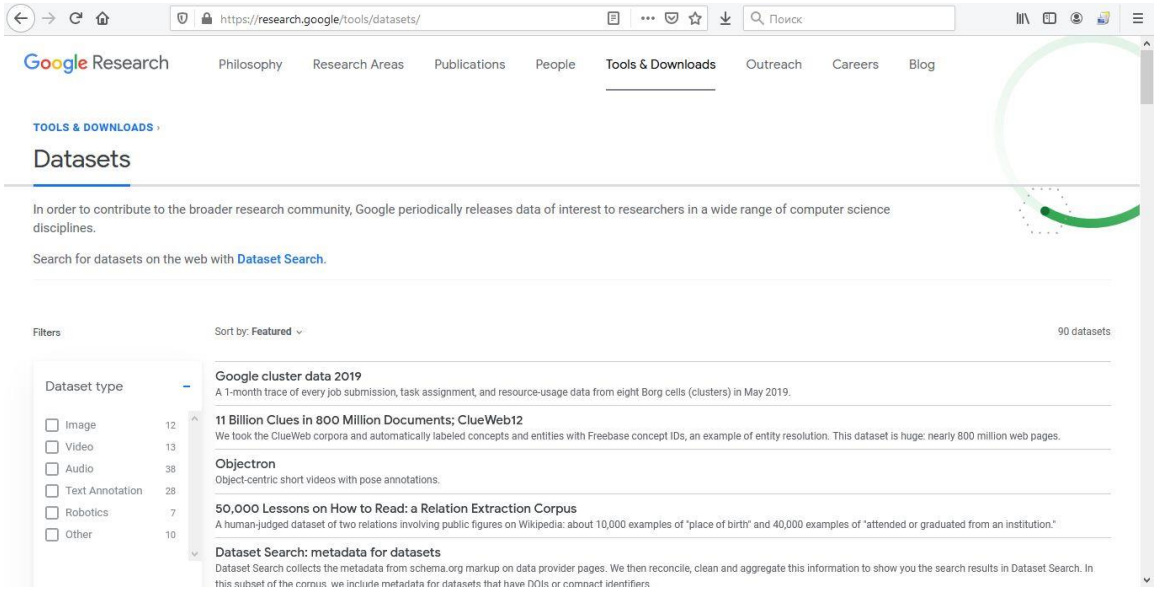

Рис. 2. Каталог наборов данных на сайте Google Research

В основном данные свободно доступны для скачивания. Однако в некоторых случаях для доступа требуется использование учетной записи Google или заполнения формы с информацией о пользователе и целях получения данных. Следует отметить, что наборы данных преимущественно характеризуются большими объемами, например, содержат большое количество файлов или же несколько файлов, размер которых исчисляется десятками и сотнями мегабайт или гигабайт.

3. Microsoft Research Open Data (https://msropendata.com/datasets). Компания Microsoft также создала репозиторий для публикации наборов данных, которые имеют отношения к её разработкам или исследованиям. Репозиторий (рис. 3) включает более 90 наборов, которые распределены по следующим категориям (в скобках приведено количество наборов на момент обращения): компьютерные науки (48), общественные науки (20), информатика (6), здравоохранение (5), физика (4), математика (3), биология (2), науки о Земле (2), образование (1), другое (1). Формат, в котором представлены данные, зависит от конкретного набора. Помимо CSV/TSV и JSON встречаются также данные в форматах DOCX, PDF, TXT и других. Часть данных доступна для просмотра непосредственно на сайте. Для скачивания файлов с данными необходимо зарегистрироваться на ресурсе.

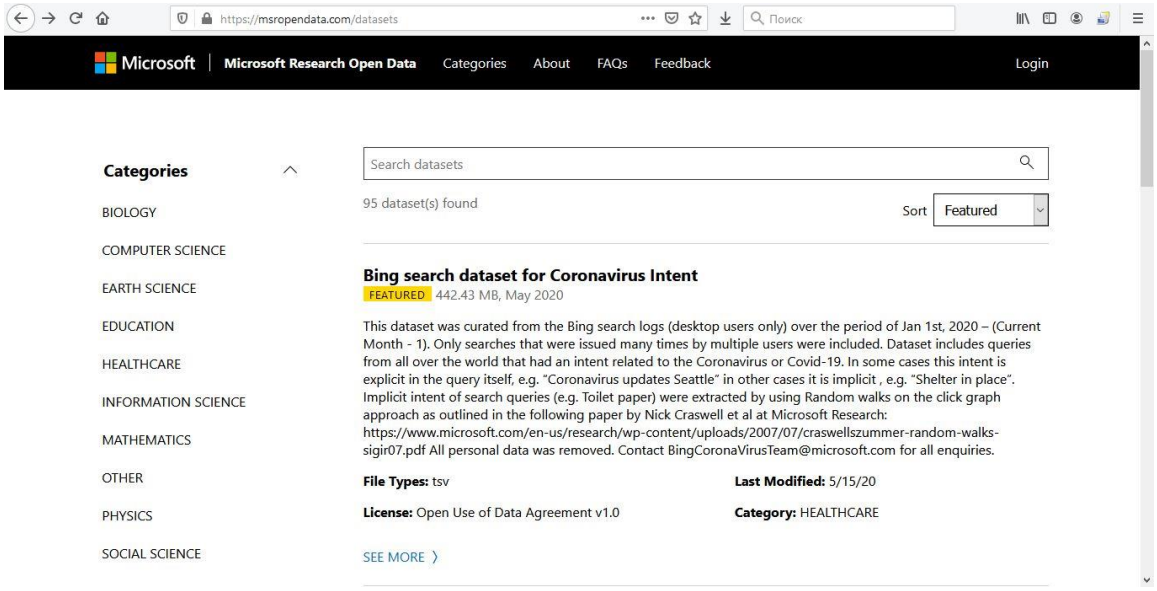

Рис. 3. Каталог наборов данных на сайте Microsoft Research Open Data

4. Registry of Open Data on AWS (https://registry.opendata.aws/). На данном ресурсе (рис. 4) можно найти более 200 наборов данных по различной тематике: от данных, относящихся к геному человека или раковым опухолям, до изображений клеточных мембран или спутниковых снимков поверхности Земли. Ключевая особенность заключается в том, что практически все наборы характеризуются большим объемом (гигабайты и терабайты), что накладывает определенные трудности при работе с ними. Данные размещаются в облачной инфраструктуре Amazon Web Services (AWS). Для каждого набора приведено описание, частота обновления, лицензия, определяющая условия использования, различные примеры использования (в том числе, связанные публикации), способ доступа к соответствующим ресурсам AWS. Для доступа и работы с данными могут использоваться различные инструменты, например, интерфейс командной строки или наборы средств разработки (SDK) для разных языков программирования и фреймворков. Для некоторых наборов данных встречаются специализированные инструменты для работы с ними.

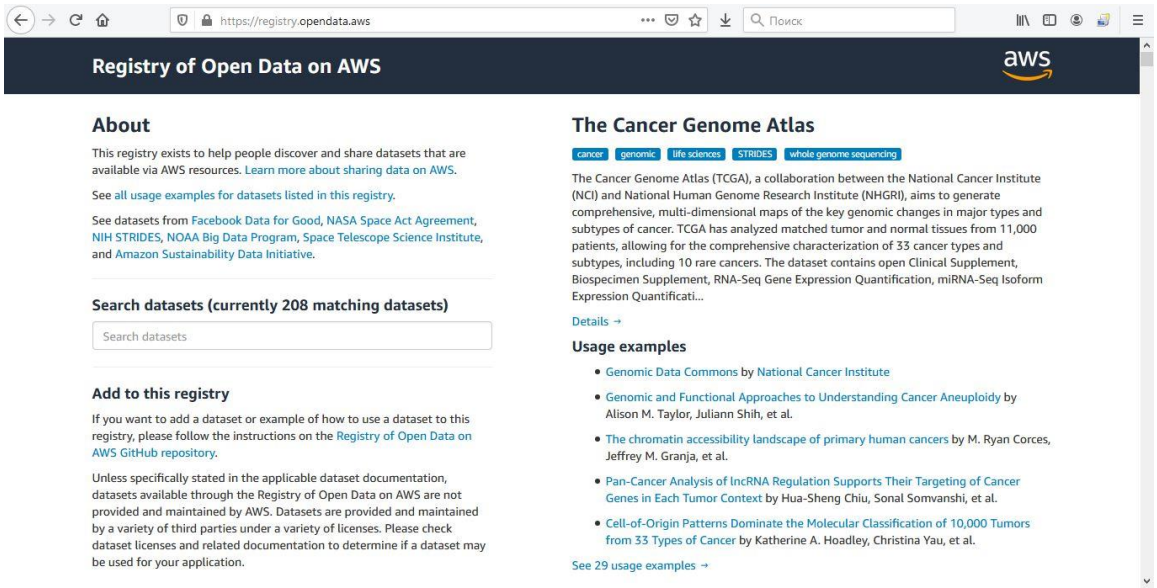

Рис. 4. Главная страница ресурса Registry of Open Data on AWS

5. Harvard Dataverse Repository (https://dataverse.harvard.edu/dataverse/harvard). Репозиторий данных Гарвардского университета включает больше 100 тысяч наборов

данных, связанных с исследованиями в различных областях науки и знаний (рис. 5). Около половины всех наборов относится к общественным наукам. Кроме того, по несколько тысяч наборов представлены для таких областей знаний, как медицина и здоровье, юриспруденция, сельскохозяйственные науки и науки о Земле. Данные доступны в самых разных форматах в зависимости от особенностей соответствующего исследования. В некоторых случаях с данными можно ознакомиться непосредственно на сайте. Любой набор доступен для свободного скачивания, как целиком, так и по частям, путем выбора только необходимых файлов.

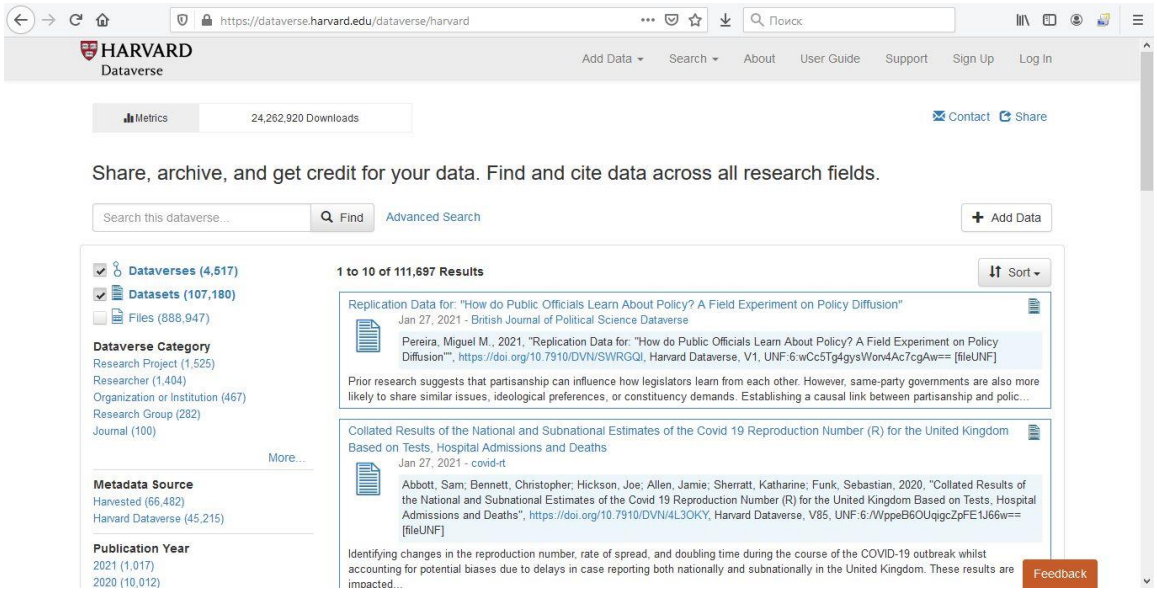

Рис. 5. Поисковая страница репозитория данных Гарвардского университета

Для каждого набора данных и составляющих его файлов сохраняется история связанных с ними изменений. В результате формируется множество версий одного и того же набора. При необходимости можно посмотреть и скачать данные для любой зафиксированной версии набора или конкретного файла из него.

Репозиторий открыт для публикации данных не только для представителей Гарвардского сообщества, но и для любых заинтересованных исследователей, желающих поделиться результатами своей работы. Для добавления набора данных требуется создать учетную запись на ресурсе.

6. Zenodo (https://zenodo.org/search?type=dataset). Данный ресурс, функционирующий под эгидой CERN, предоставляет возможность публиковать научные работы и результаты исследований в различной форме (статьи, книги, презентации, изображения, программное обеспечение и др.) [30]. В том числе на сайте доступно более 65 тысяч наборов данных (рис. 6). В основном представленные наборы доступны для свободного ознакомления и скачивания. В некоторых случаях данные могут быть просмотрены непосредственно на сайте. Каждый набор данных может иметь несколько отличающихся версий. Ресурс предоставляет средства для просмотра списка версий, а также выбранной версии набора данных. Однако эти средства не настолько функциональны, как в репозитории данных Гарвардского университета. Кроме того, существует программный интерфейс (REST API) для работы с ресурсом (например, для поиска, скачивания и загрузки данных).

Для публикации результатов исследований (например, нового набора данных) или использования программного интерфейса требуется регистрация на ресурсе. К недостаткам ресурса можно отнести то, что отсутствует возможность скачать сразу целиком все файлы,

связанные с конкретным набором данных. Каждый файл необходимо скачивать по отдельности.

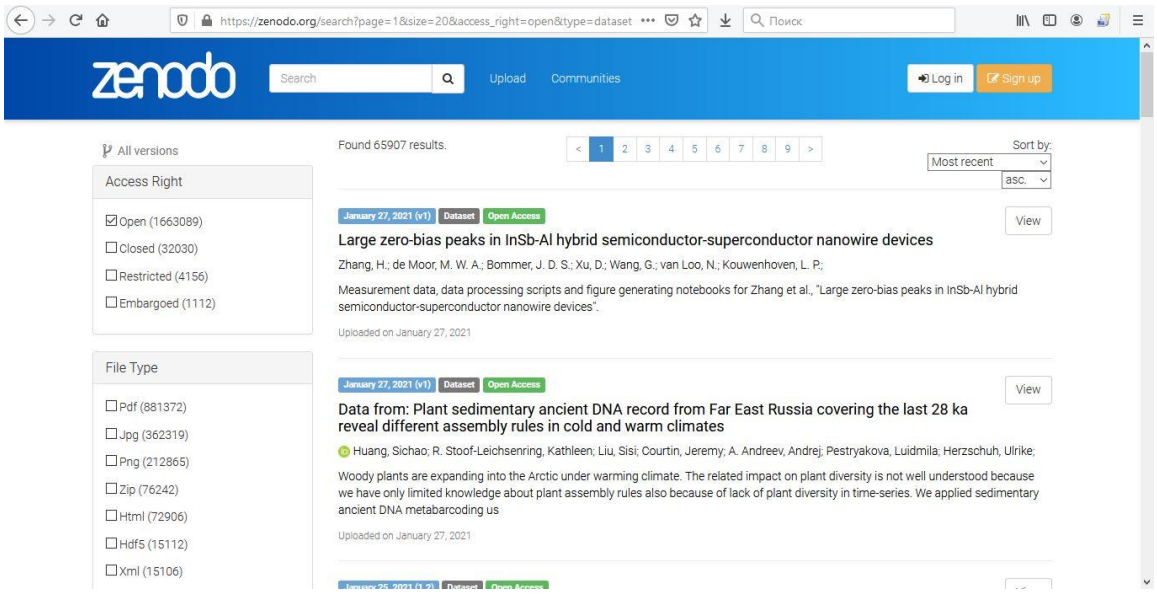

Рис. 6. Каталог наборов данных на сайте Zenodo

7. Портал открытых данных Российской Федерации (https://data.gov.ru/opendata). Реестр на портале открытых данных Российской Федерации включает более 20 тысяч наборов данных [25], классифицированных по различным рубрикам (рис. 7). Преимущественно публикуются данные государственных органов власти, субъектов федерации или муниципальных органов власти. В основном данные относятся к категории «Государство» (15152 набора на момент обращения). Однако имеется много наборов, входящих в такие рубрики, как «Экономика» (1552 набора на момент обращения), «Образование» (1242), «Экология» (1229), «Здоровье» (673), «Культура» (630), «Транспорт» (625) и другие (всего представлено 16 рубрик). На портале можно найти, например, такие разнообразные данные, как «Единый реестр субъектов малого и среднего предпринимательства», «Государственный реестр лекарственных средств», «Государственный реестр сертифицированных средств защиты информации» или «Перечень стран и режимов въезда на их территорию». Данные представлены в различных форматах, но превалируют CSV (больше 15 тысяч наборов на момент обращения), XML и JSON. Для некоторых наборов доступны средства просмотра данных и их структуры непосредственно на портале. Есть возможность скачать в виде файла формата Excel или CSV информацию о наборах, входящих в реестр. Кроме того, портал предоставляет программный интерфейс для работы с данными, с помощью которого можно, например, получить сведения о том или ином наборе данных или загрузить его.

Для поиска, просмотра и скачивания данных не требуется регистрация на портале. Регистрация может понадобиться для пользования расширенными функциями портала, например для подачи заявки для публикации нового набора данных или применения программного интерфейса.

8. World Bank (https://datacatalog.worldbank.org https://datatopics.worldbank.org/world-development-indicators/). Сайт Всемирного банка предоставляет доступ к большому количеству экономических и статистических данных (рис. 8) [25]. Представлены несколько тысяч наборов данных, преимущественно касающихся различных показателей развития стран и их экономик, демографии, окружающей среды, а также отчеты и результаты исследований Всемирного банка. Например, можно найти гендерную статистику

относительно демографии, здоровья, экономического благополучия, а также различные экономические показатели типа валового национального дохода стран. Данные доступны в формате Excel или CSV. Кроме того, на сайте Всемирного банка представлен инструмент (https://databank.worldbank.org), позволяющий выполнять анализ и визуализацию на основе доступных данных. В частности, инструмент позволяет делать выборки и строить графики по различным показателям, странам и годам.

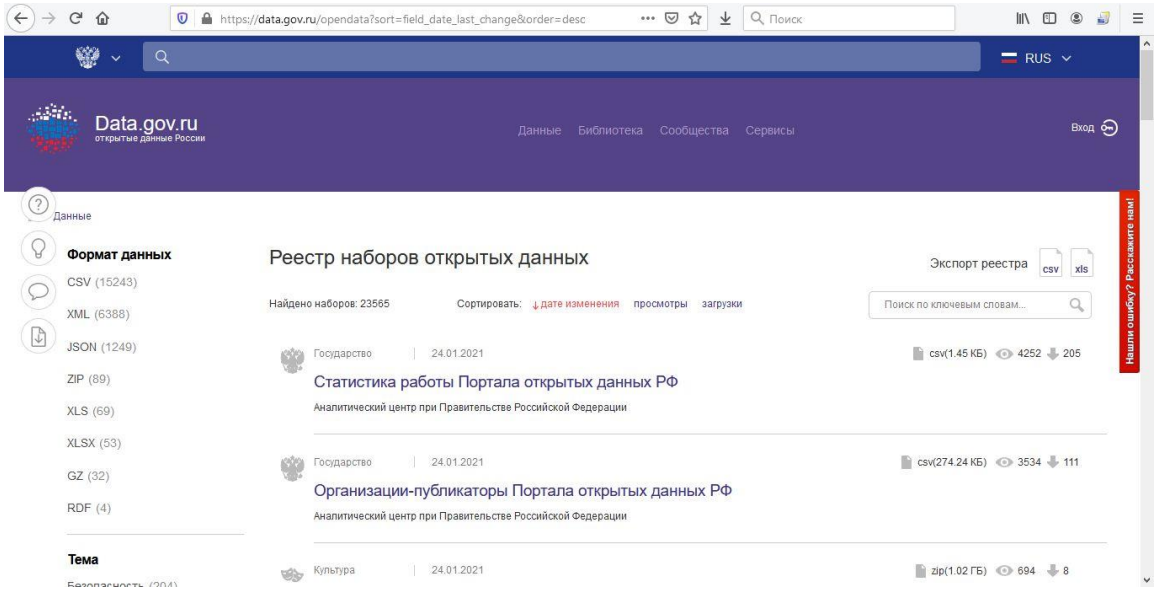

Рис. 7. Реестр наборов данных на Портале открытых данных России

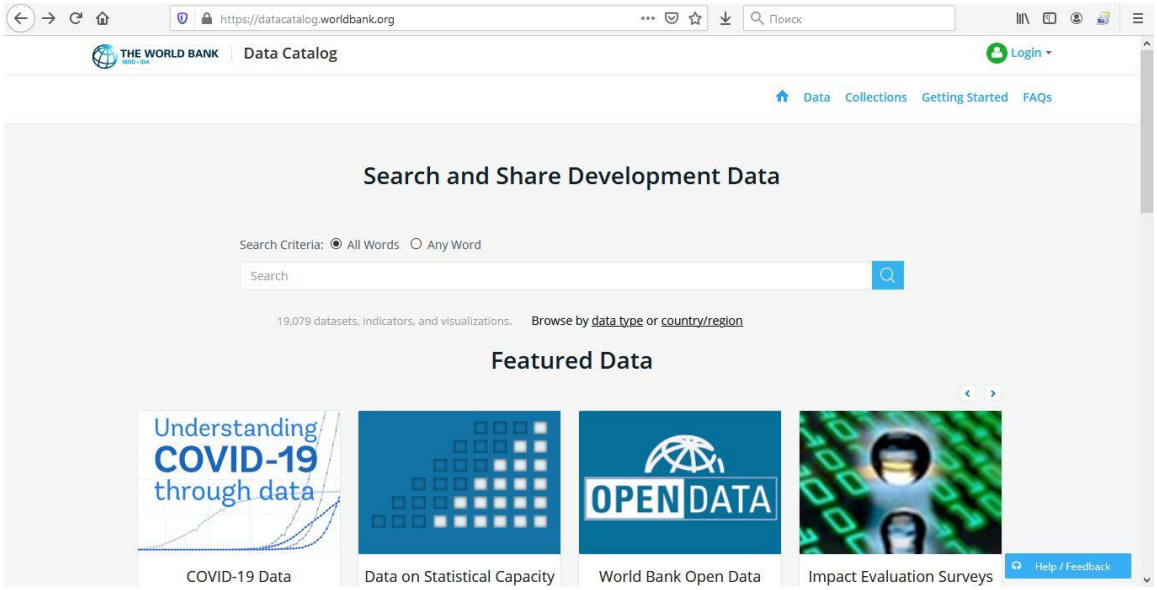

Рис. 8. Главная страница каталога данных на сайте Всемирного банка

9. The Big Bad NLP Database (https://datasets.quantumstat.com). Каталог ссылок на ресурсы, на которых можно найти наборы данных для разнообразных задач, относящихся к области обработки языка (преимущественно естественного). На сайте представлены более 800 ресурсов (рис. 9). Для каждого набора данных приведено описание, дата добавления в каталог, язык, объем данных (если известен), формат файлов с данными (если известен), категории целевых задач (например, выявление эмоциональной оценки, классификация, извлечение информации и т.п.), год формирования набора и ссылка на сайт, на котором можно непосредственно найти соответствующие данные. Большинство элементов каталога снабжены ссылками на сопутствующие публикации, в которых рассматривается тот или

иной набор данных. В основном наборы данных относятся к английскому языку (более 450 на момент посещения каталога) или сразу к нескольким языкам (более 80). Однако в каталоге встречаются наборы более чем для 100 разных языков, в том числе для китайского (21 на момент посещения каталога), арабского (19), индонезийского (14) и русского (12). Есть наборы для двух языков, например немецкий-английский (7 на момент посещения каталога). Кроме ресурсов, относящихся к естественным языкам, в каталоге представлены несколько наборов данных, ориентированных на задачи, связанные с обработкой языков программирования (например, выявление семантической эквивалентности или безопасности фрагментов кода).

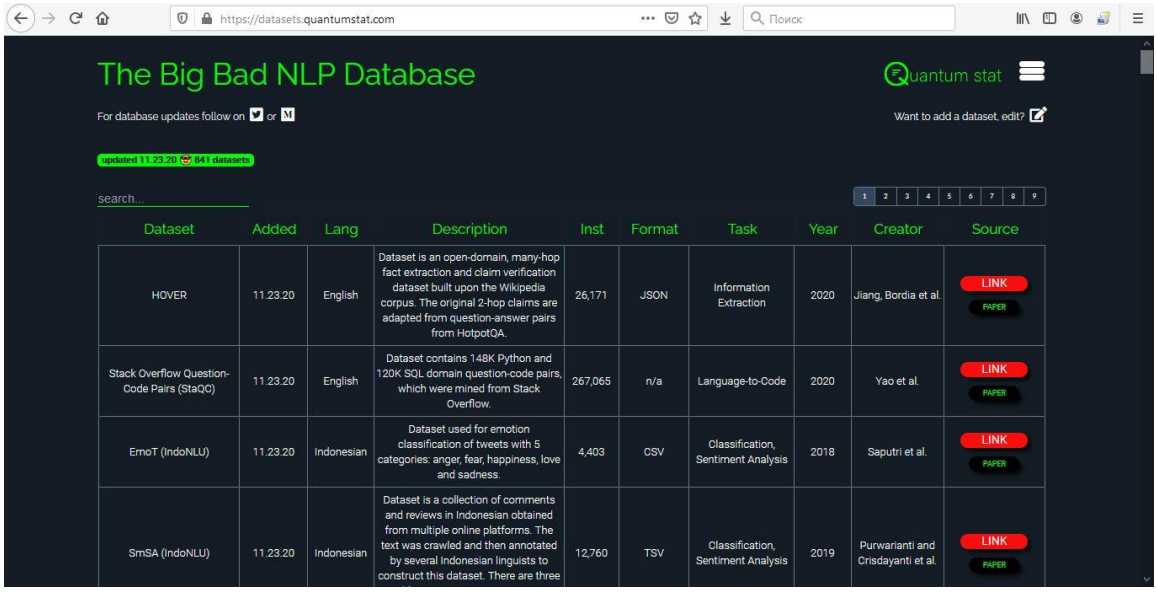

Рис. 9. Главная страница каталога The Big Bad NLP Database

10. Common Crawl (https://commoncrawl.org https://index.commoncrawl.org). Данный ресурс обеспечивает доступ к копиям web-страниц сети Интернет, которые автоматически собираются на регулярной основе, начиная с 2008 года. Набор содержит миллиарды страниц, представленных на различных сайтах на момент сбора данных. Кроме непосредственно самих сохраненных web-страниц, доступны связанные с ними метаданные, а также текст, извлеченный из страниц. В связи с этим данный набор может использоваться, в частности, для различных задач, связанных с обработкой естественного языка или поиска неструктурированной информации, представленной в Интернет. Собранные данные хранятся в специальном формате, описание которого приведено на сайте ресурса. Для работы с данными могут использоваться различные средства и инструменты (например, на базе Java или Python), которые перечислены на отдельной странице ресурса (рис. 10). Также представлен ряд обучающих материалов, поясняющих и демонстрирующих различные аспекты по работе с набором данных. Кроме того, есть отдельный сайт Common Crawl Index Server (https://index.commoncrawl.org), с помощью которого можно получить быстрый доступ к сведениям, относящимся к тому или иному домену или URL. Например, с его помощью можно проверить наличие страницы с определенным адресом в наборе данных.

Заключение. Рассмотренные ресурсы различаются по таким характеристикам, как количество и объем представленных наборов данных, предметные области и прикладные задачи, к которым относятся данные, удобство и особенности доступа к ним и др. В совокупности с сайтами, перечисленными в [27], данное множество ресурсов способно в значительной мере покрыть потребность в поиске данных, требуемых отладки методов и инструментария в разных сферах применения искусственного интеллекта и машинного

обучения. В том числе, указанные ресурсы могут быть полезны для нахождения данных с целью проверки научных гипотез или использования в процессе обучения. Дополнительно для поиска подходящих наборов данных можно воспользоваться следующими сайтами: Data Portals (http://dataportals.org), Dataset Search (https://datasetsearch.research.google.com), KEEL repository (https://sci2s.ugr.es/keel/datasets.php), Open Datasets (https://wiki.pathmind.com/open-datasets).

В работе [25] перечислены несколько ресурсов, на которых представлены социальные и экономические данные, относящиеся к Российской Федерации.

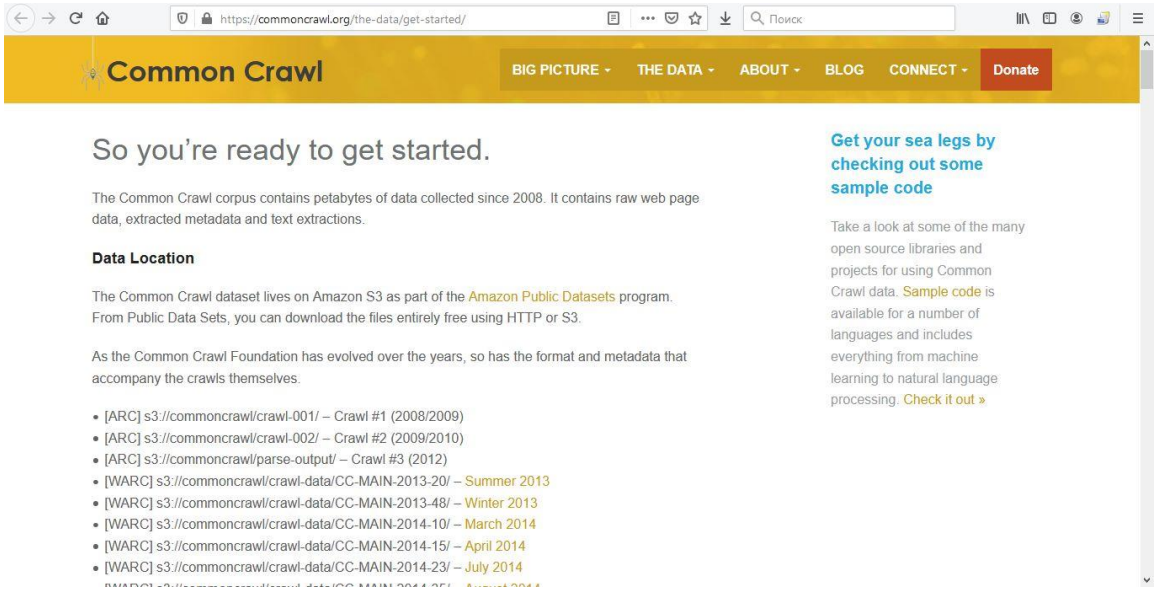

Рис. 10. Страница ресурса Common Crawl, предоставляющая доступ к данным и средствам для работы с ними

СПИСОК ЛИТЕРАТУРЫ

1. Брюхина Н.Г., Рева П.В., Баранников В.А. Интеллектуальная автоматизация как драйвер экономики в условиях пандемии // Ресурсосбережение. Эффективность. Развитие. Матер. V республиканской науч.-практ. конф. 2020. С. 443-450.
2. Воробьева А.В. Поствирусные тенденции: отраслевые изменения // Инновации. Наука. Образование. 2020. № 16. С. 584-589.
3. Цогоева М.И., Галаова Э.О. Влияние пандемии covid-19 на использование информационных технологий в международном бизнесе // Актуальные вопросы современной экономики. 2020. № 5. С. 418-427. DOI:10.34755/IROK.2020.42.71.168.
4. Ашинов К.В. Цифровые технологии в условиях пандемии коронавируса // Проблемы и перспективы развития России: молодежный взгляд в будущее. Сб. науч. ст. 3-й Всероссийской науч. конф. Курск. 2020. С. 137-139.
5. Гусев А.В., Новицкий Р.Э. Технологии прогнозной аналитики в борьбе с пандемией COVID-19 // Врач и информационные технологии. 2020. № 4. С. 24-33. DOI:10.37690/1811-0193-2020-4-24-33.
6. Евсюков В.В., Свиридова Т.В., Богатенко Е.Р. Искусственный интеллект и коронавирус COVID-19 // Вестник Тульского филиала финуниверситета. 2020. № 1. С. 295-297.
7. Иванов М.В., Румянцева С.Ю. Новая экосистема цифровой недвижимости: этапы развития, технологии и перспективы // Известия вузов. Инвестиции. Строительство. Недвижимость. 2020. Т. 10. № 4. С. 524-533. DOI:10.21285/2227-2917-2020-4-524-533.

8. Козырева О.Н., Ольхова Л.А. Особенности применения дистанционного обучения в условиях пандемии // Общество, педагогика, психология. Сб. матер. Всероссийской науч.-практ. конф. 2020. С. 37-41.
9. Михайлов А.А., Федулов В.И. Подходы к управлению персоналом в условиях удаленной занятости // Естественно-гуманитарные исследования. 2020. № 29(3). С. 222-225. DOI: 10.24411/2309-4788-2020-10262.
10. Сингатулин В.Н., Дудаков Г.С. Влияние COVID-19 на цифровизацию в пищевой промышленности // Приоритеты экономического роста страны и регионов в период постпандемии. Сб. матер. Всероссийской науч.-практ. конф. 2020. С. 274-278.
11. Стародубов И.И., Данилов И.А. Тренды в банковском финтехе 2021 года // Матрица научного познания. 2021. № 2-1. С. 114-120.
12. Шалина Д.С., Степанова Н.Р. Теория и практика использования искусственного интеллекта в сфере недвижимости // Вестник Алтайской академии экономики и права. 2020. № 5-1. С. 193-200. DOI:10.17513/vaael.1128.
13. Качнов С.А. Методы машинного обучения в медицине // Наука молодых - будущее России. Сб. науч. ст. 5-й Межд. науч. конф. перспективных разработок молодых ученых. Курск. 2020. С. 47-50.
14. Самбурский С.Е., Сергунова К.А. Московский эксперимент по компьютерному зрению в лучевой диагностике // Московская медицина. 2020. № 4(38). С. 32-39.
15. Ярмухаметов Р.Р. Обзор применений искусственного интеллекта в медицине // Наукосфера. 2020. № 12-2. С. 172-178.
16. Иванникова В.П., Шелухин О.И. Бинарная классификация компьютерных атак на примере базы данных UNSW-NB15 // Телекоммуникации и информационные технологии. 2020. № 1. С. 10-18.
17. Лебедев Г.С., Маслюков А.П.., Шадеркин И.А., Шадеркина А.И. Глубокое машинное обучение (искусственный интеллект) в ультразвуковой диагностике // Журнал телемедицины и электронного здравоохранения. 2020. № 2. С. 22-29. DOI:10.29188/2542-2413-2020-6-2-22-29.
18. Марахтанов А.Г., Паренченков Е.О., Смирнов Н.В. Определение электронного мошенничества методами машинного обучения в случае несбалансированного набора данных // Вестник пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2020. №36. С. 80-95.
19. Рунова К.В., Юрин А.А. Классификация сердечно-сосудистых заболеваний с помощью инструментальных методов обработки информации на основе различных методов машинного обучения // Colloquium-journal. 2019. № 13-3(37). С. 115-120.
20. Фомин В.В., Александров И.В. Об одном опыте применения web - инструментария машинного обучения // Моделирование и анализ сложных технических и технологических систем: сборник статей по итогам Международной научно-практической конференции. Самара. 2018. С. 131-137.
21. Saif M.A., Medvedev A.N., Medvedev M.A., Atanasova T. Classification of online toxic comments using the logistic regression and neural networks models // AIP Conference Proceedings 2048, 060011 (2018). DOI:10.1063/1.5082126.
22. Shtovba S., Shtovba O., Yahymovych O., Petrychko M. Impact of the syntactic dependencies in the sentences on the quality of the identification of the toxic comments in the social networks // Scientific works of Vinnytsia national technical university. 2019. № 4. Pp. 35-42. DOI: 10.31649/2307-5392-2019-4-35-42.

23. Аверина М.Д. Применение сверточных нейронных сетей в задаче классификации медицинских изображений // Заметки по информатике и математике. 2019. Вып. 11. С. 1-9.
24. Венцов Н.Н., Подколзина Л.А. Общий подход к созданию набора данных на примере формирования набора изображений линейных штрих-кодов // Journal of advanced research in technical science. 2020. № 18. С. 50-54. DOI:10.26160/2474-5901-2020-18-50-54.
25. Kashirina A.M., Kravchenko A.V. Problems of open data sources analysis for socio-economic and medical research // The European Proceedings of Social and Behavioural Sciences. Krasnoyarsk. 2020. Pp. 1604-1612. DOI:10.15405/epsbs.2020.10.03.184.
26. Tymchenko B., Marchenko Ph., Spodarets D. Segmentation of cloud organization patterns from satellite images using deep neural networks // Herald of Advanced Information Technology. 2020. Vol.3. № 1. Pp. 352-361.
27. Сикулер Д.В. Поиск данных для апробации интеллектуальных алгоритмов и технологий // Межд. науч. журнал «Символ науки». 2020. № 4. С. 49–54.
28. Багаев И.В., Коломенская И.Д., Шатров А.В. Алгоритм наивного метода Байеса в задачах бинарной классификации на примере набора данных Santander с платформы Kaggle // Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века: сб. ст. по материалам Четвертой всерос. науч.-практ. конф. Пермь. 2019. С. 32-36.
29. Кесян Г.Р., Воронова Л.И., Трунов А.С. Прогнозирование наличия сахарного диабета с использованием нейронных сетей // Технологии информационного общества. Материалы XIII Международной отраслевой научно-технической конференции. 2019. С. 438-440.
30. Чадин И.Ф. Zenodo и GBIF: инструменты для публикации наборов первичных данных // Вестник Института биологии Коми НЦ УрО РАН. 2018. № 3(205). С. 34-36. DOI: 10.31140/j.vestnikib.2018.3(205).5.

UDK 004.8+004.62

RESOURCES PROVIDING DATA FOR MACHINE LEARNING AND TESTING ARTIFICIAL INTELLIGENCE TECHNOLOGIES

Denis V. Sikuler

Candidate of Technical Sciences, Associate Professor of Information systems Department,
e-mail: rimol@km.ru,
Herzen State Pedagogical University of Russia,
191186, Russia, St. Petersburg, 48 Moika Embankment.

Annotation. The work presents review of 10 Internet resources that can be used to find data for different tasks related to machine learning and artificial intelligence. There were examined some popular sites (like Kaggle, Registry of Open Data on AWS) and some less known and specific ones (like The Big Bad NLP Database, Common Crawl). All included resources provide free access to data. Moreover in most cases registration is not needed for data access. Main features are specified for every examined resource, including regarding data search and access. The following sites are included in the review: Kaggle, Google Research, Microsoft Research Open Data, Registry of Open Data on AWS, Harvard Dataverse Repository, Zenodo, Open Data portal of the Russian Federation, World Bank, The Big Bad NLP Database, Common Crawl.

Keywords: data search, dataset, data set, open data, data repository, dataset catalog, artificial intelligence, machine learning.

REFERENCES

1. Bryukhina N.G., Reva P.V., Barannikov V.A. Intellektual'naya avtomatizatsiya kak drayver ekonomiki v usloviyakh pandemii [Intelligent automation as a driver of the economy in a pandemic] // Resursosberezhenie. Effektivnost'. Razvitie. Mater. V respublikanskoy nauch.-prakt. konf. = Resource saving. Efficiency. Development. Proceedings of V republican scientific and practical conf. 2020. Pp. 443-450.
2. Vorob'eva A.V. Postvirusnye tendentsii: otraslevye izmeneniya [Post-viral trends: industry changes] // Innovatsii. Nauka. Obrazovanie = Innovation. Science. Education. 2020. № 16. Pp. 584-589.
3. Tsogoeva M.I., Galaova E.O. Vliyanie pandemii covid-19 na ispol'zovanie informatsionnykh tekhnologiy v mezhdunarodnom biznese [Impact of the covid-19 pandemic on the use of information technology in international business] // Aktual'nye voprosy sovremennoy ekonomiki = Topical issues of the modern economy. 2020. № 5. Pp. 418-427. DOI:10.34755/IROK.2020.42.71.168.
4. Ashinov K.V. Tsifrovye tekhnologii v usloviyakh pandemii koronavirusa [Digital technologies in the context of the coronavirus pandemic] // Problemy i perspektivy razvitiya Rossii: molodezhnyy vzglyad v budushchee. Sb. nauch. st. 3-y Vserossiyskoy nauch. konf. Kursk = Problems and Prospects for the Development of Russia: a Youth Look into the Future. Proceedings of the 3rd Russian scientific conf. Kursk. 2020. Pp. 137-139.
5. Gusev A.V., Novitskiy R.E. Tekhnologii prognoznoy analitiki v bor'be s pandemiey COVID-19 [Predictive analytics technologies in the management of the COVID-19 pandemic] // Vrach i informatsionnye tekhnologii = Information technologies for the Physician. 2020. № 4. Pp. 24-33. DOI: 10.37690/1811-0193-2020-4-24-33.
6. Evsyukov V.V., Sviridova T.V., Bogatenko E.R. Iskusstvennym intellekt i koronavirus COVID-19 [Artificial intelligence and COVID-19 coronavirus] // Vestnik Tul'skogo filiala finuniversiteta = Bulletin of the Tula branch of the financial university. 2020. № 1. Pp. 295-297.
7. Ivanov M.V., Rumyantseva S.Yu. Novaya ekosistema tsifrovoy nedvizhimosti: etapy razvitiya, tekhnologii i perspektivy [A new ecosystem of digital real estate: Developmental stages, technologiesand prospects] // Izvestiya vuzov. Investitsii. Stroitel'stvo. Nedvizhimost' = Proceedings of Universities. Investment. Construction. Realestate. 2020. 10(4). Pp. 524-533. DOI:10.21285/2227-2917-2020-4-524-533.
8. Kozyreva O.N., Ol'khova L.A. Osobennosti primeneniya distantsionnogo obucheniya v usloviyakh pandemii [Specifics of the use of distance learning in a pandemic] // Obshchestvo, pedagogika, psikhologiya. Sb. mater. Vserossiyskoy nauch.-prakt. konf. = Society, pedagogy, psychology. Proceedings of Russian scientific and practical conf. 2020. Pp. 37-41.
9. Mikhaylov A.A., Fedulov V.I. Podkhody k upravleniyu personalom v usloviyakh udalennoy zanyatosti [Approaches to human resources management in remote employment] // Estestvenno-gumanitarnye issledovaniya = Natural humanitarian studies. 2020. № 29(3). Pp. 222-225. DOI:10.24411/2309-4788-2020-10262.
10. Singatulin V.N., Dudakov G.S. Vliyanie COVID-19 na tsifrovizatsiyu v pishchevoy promyshlennosti [Impact of COVID-19 on digitalization in the food industry] // Prioritety

ekonomicheskogo rosta strany i regionov v period postpandemii. Sb. mater. Vserossiyskoy nauch.-prakt. konf. = Priorities for economic growth of the country and regions in the post-pandemic period. Proceedings of Russian scientific and practical conf. 2020. Pp. 274-278.

11. Starodubov I.I., Danilov I.A. Trendy v bankovskom fintekhe 2021 goda [Trends in banking fintech 2021] // Matritsa nauchnogo poznaniya. = Matrix of Scientific Cognition. 2021. № 2-1. Pp. 114-120.
12. Shalina D.S., Stepanova N.R. Teoriya i praktika ispol'zovaniya iskusstvennogo intellekta v sfere nedvizhimosti [Theory and practice of using artificial intelligence in real estate] // Vestnik Altayskoy akademii ekonomiki i prava = Bulletin of the Altai Academy of Economics and Law. 2020. № 5-1.
13. Kachnov S.A. Metody mashinnogo obucheniya v meditsine [Machine learning methods in medicine] // Nauka molodykh - budushchee Rossii. Sb. nauch. st. 5-y Mezhd. nauch. konf. perspektivnykh razrabotok molodykh uchenykh. Kursk = The science of the young is the future of Russia. Proceedings of 5th Int. scientific conf. of promising developments of young scientists. Kursk. 2020. Pp. 47-50.
14. Samburskiy S.E., Sergunova K.A. Moskovskiy eksperiment po komp'yuternomu zreniyu v luchevoy diagnostike [Moscow Experiment with Using a Computer Vision in Diagnostic Radiology] // Moskovskaya meditsina = Moscow medicine. 2020. № 4(38). Pp. 32-39.
15. Yarmukhametov R.R. Obzor primeneniy iskusstvennogo intellekta v meditsine [Overview of usages of artificial intelligence in medicine] // Naukosfera = Naukosfera. 2020. № 12-2. Pp. 172-178.
16. Ivannikova V.P., Shelukhin O.I. Binarnaya klassifikatsiya komp'yuternykh atak na primere bazy dannykh UNSW-NB15 [Computer attacks binary classification on the UNSW-NB15 dataset example] // Telekommunikatsii i informatsionnye tekhnologii = Telecommunications and information technologies. 2020. № 1. Pp. 10-18.
17. Lebedev G.S., Maslyukov A.P.., Shaderkin I.A., Shaderkina A.I. Glubokoe mashinnoe obuchenie (iskusstvennyy intellekt) v ul'trazvukovoy diagnostike [Deep machine learning (artificial intelligence) in ultrasound diagnostics] // Zhurnal telemeditsiny i elektronnogo zdravookhraneniya = Journal of Telemedicine and E-Health. 2020. № 2. Pp. 22-29. DOI: 10.29188/2542-2413-2020-6-2-22-29.
18. Marakhtanov A.G., Parenchenkov E.O., Smirnov N.V. Opredelenie elektronnogo moshennichestva metodami mashinnogo obucheniya v sluchae nesbalansirovannogo nabora dannykh [Fraud detection by machine learning methods in the case of an imbalanced dataset] // Vestnik permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniya = PNRPU Bulletin. Electrotechnics, Informational Technologies, Control Systems. 2020. № 36. Pp. 80-95.
19. Runova K.V., Yurin A.A. Klassifikatsiya serdechno-sosudistykh zabolevaniy s pomoshch'yu instrumental'nykh metodov obrabotki informatsii na osnove razlichnykh metodov mashinnogo obucheniya [Classification of cardiovascular diseases using instrumental information processing methods based on various machine learning methods] // Colloquium-journal = Colloquium-journal. 2019. № 13-3 (37). Pp. 115-120.
20. Fomin V.V., Aleksandrov I.V. Ob odnom opyte primeneniya web - instrumentariya mashinnogo obucheniya [About an experience of using web-based machine learning tools] // Modelirovanie i analiz slozhnykh tekhnicheskikh i tekhnologicheskikh sistem: sbornik statey po itogam Mezhdunarodnoy nauchno-prakticheskoy konferentsii. Samara = Modeling

and analysis of complex technical and technological systems: a collection of articles on the results of the International scientific and practical conference. Samara. 2018. Pp. 131-137.

21. Saif M.A., Medvedev A.N., Medvedev M.A., Atanasova T. Classification of online toxic comments using the logistic regression and neural networks models // AIP Conference Proceedings 2048, 060011 (2018). DOI:10.1063/1.5082126.
22. Shtovba S., Shtovba O., Yahymovych O., Petrychko M. Impact of the syntactic dependencies in the sentences on the quality of the identification of the toxic comments in the social networks // Scientific works of Vinnytsia national technical university. 2019. № 4. Pp. 35-42. DOI:10.31649/2307-5392-2019-4-35-42.
23. Averina M.D. Primenenie svertochnykh neyronnykh setey v zadache klassifikatsii meditsinskikh izobrazheniy [Application of convolutional neural networks in the problem of classification of medical images] // Zametki po informatike i matematike = Notes on computer science and mathematics. 2019. Issue 11. Pp. 1-9.
24. Ventsov N.N., Podkolzina L.A. Obshchiy podkhod k sozdaniyu nabora dannykh na primere formirovaniya nabora izobrazheniy lineynykh shtrikh-kodov [The general approach to creating a dataset using an example of barcode images] // Journal of advanced research in technical science. 2020. № 18. Pp. 50-54. DOI:10.26160/2474-5901-2020-18-50-54.
25. Kashirina A.M., Kravchenko A.V. Problems of open data sources analysis for socio-economic and medical research // The European Proceedings of Social and Behavioural Sciences. Krasnoyarsk. 2020. Pp. 1604-1612. DOI:10.15405/epsbs.2020.10.03.184.
26. Tymchenko B., Marchenko Ph., Spodarets D. Segmentation of cloud organization patterns from satellite images using deep neural networks // Herald of Advanced Information Technology. 2020. Vol.3. № 1. Pp. 352-361.
27. Sikuler D.V. Poisk dannykh dlya aprobatsii intellektual'nykh algoritmov i tekhnologiy [Search of data to test intellectual algorithms and technologies] // Mezhd. nauch. zhurnal “Simvol nauki” = International scientific journal “Symbol of science”. 2020. № 4. Pp. 49–54.
28. Bagaev I.V., Kolomenskaya I.D., Shatrov A.V. Algoritm naivnogo metoda Bayesa v zadachakh binarnoy klassifikatsii na primere nabora dannykh Santander s platformy Kaggle [Algorithm of naive Bayes methods in binary classification tasks on Santander dataset example from Kaggle platform] // Iskusstvennyy intellekt v reshenii aktual'nykh sotsial'nykh i ekonomicheskikh problem XXI veka: sb. st. po materialam Chetvertoy vseros. nauch.-prakt. konf. Perm' = Artificial intelligence in solving urgent social and economic problems of the XXI century: proceedings of the 4th Russia scientific and practical conf. Perm. 2019. Pp. 32-36.
29. Kesyan G.R., Voronova L.I., Trunov A.S. Prognozirovanie nalichiya sakharnogo diabeta s ispol'zovaniem neyronnykh setey [Predicting the presence of diabetes mellitus using neural networks] // Tekhnologii informatsionnogo obshchestva. Materialy XIII Mezhdunarodnoy otraslevoy nauchno-tekhnicheskoy konferentsii = Information Society Technologies. Proceedings of the XIII International branch scientific and technical conference. 2019. Pp. 438-440.
30. Chadin I.F. Zenodo i GBIF: instrumenty dlya publikatsii naborov pervichnykh dannykh [Zenodo and GBIF: tools for scientific primary data publication] // Vestnik Instituta biologii Komi NTs UrO RAN = Bulletin of the Institute of Biology of Komi Science Centre of the Ural Branch of the Russian Academy of Sciences. 2018. № 3(205). Pp. 34-36. DOI: 10.31140/j.vestnikib.2018.3(205).5.