24 открытых набора данных для ваших проектов Data Science/ML

Описание

Поиск нужных наборов данных может оказаться сложной задачей, особенно если они нужны для проектов в области машинного обучения (ML) и науки о данных. Мы сокращаем ваши усилия по поиску, предоставляя полный список бесплатных наборов данных. Наборы данных – это просто коллекции данных. Это могут быть финансовые данные, данные о здоровье населения, данные фондового рынка, банковские данные, географические данные, данные исследований в области науки о частицах, рейтинги товаров на сайте электронной коммерции и т. д. Наборы данных содержат сведения, собранные по стандарту научного исследования, и важны для дальнейшей визуализации, извлечения, прогнозирования и т. д. Поскольку данные – это эквивалент сырой нефти в цифровой вселенной, наборы данных становятся коммерческими и дефицитными. Продолжайте читать, чтобы узнать об основах работы с наборами данных. Вы также узнаете о некоторых открытых наборах данных, которые действительно бесплатны для ваших проектов в области машинного обучения (ML) или науки о данных.

Что такое наборы данных?

Наборы данных – это совокупность данных в структурированном и организованном виде. Обычно исследователи связывают наборы данных с уникальным органом, например, World Bank Open Data. Опять же, сборщики данных хранят наборыданных по конкретной теме, как, например, данные переписи населенияСоединенных Штатов Америки 2020 года, опубликованные Бюро переписинаселения США.

Вы найдете множество наборов данных по глобальным и местным проблемам. Большинство наборов данных содержат взаимосвязанные точки данных. Например, население страны и то, как ожирение относится к различным классам этого населения. Специалистам по изучению данных может потребоваться очистить, реструктурировать и обработать такие массивы данных с помощью инструментов больших данных, чтобы прийти к ценным выводам, например, сократить количество пластиковых отходов, проанализировав данные об использовании пластика, решить проблемы с рабочей силой, проанализировав данные о заработной плате, обучить искусственный интеллект (ИИ) и т. д.

Типы наборов данных

В зависимости от источника данных они могут быть государственными или частными. Публичные наборы данных открыты для всех и вносят большой вклад в исследования и разработки.

В зависимости от содержащейся в них информации наборы данных могут быть следующих типов:

- **Многомерные:** Такие данные содержат множество переменных.
- **Категоричность:** изображает множество категорий людей.
- **Числовые:** такие наборы данных измеряют данные в цифрах, например, возраст, рост и т.д.
- **Корреляция:** В этом типе точки данных взаимосвязаны.
- **Файловая система:** Здесь наборы данных хранятся в файлах.
- **Двумерный:** Набор данных, содержащий две переменные и взаимосвязь между ними.
- **Набор веб-данных:** Данные, собранные с одного или многих аналогичных интернет-порталов.
- **База данных:** В таких наборах данных данные хранятся в виде таблиц, столбцов и строк.

Наборы данных с открытым исходным кодом для проектов в области науки о данных

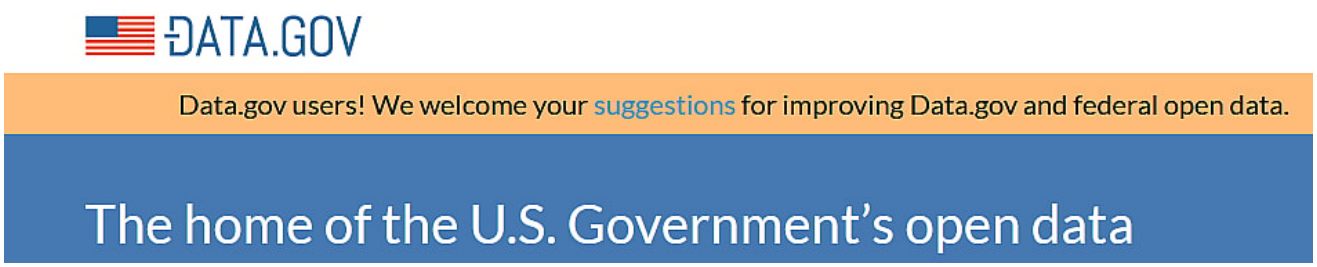
Бесплатные наборы данных – это топливо для вашей страсти к карьере в области науки о данных. Поскольку если вы находитесь на ранних этапах карьеры в области науки о данных, вы можете захотеть заняться личными и некоммерческими проектами для уверенности в себе или создания портфолио.

Во-первых, вы сможете легко проверить полученные навыки, применяя инструменты и методы для решения реальных задач с наборами данных. Например, в свободном доступе находятся данные исследований рака, данные Covid-19, данные ФБР о криминальных сводках, данные анализа частиц из ЦЕРН и т. д. Вы можете использовать такие данные и построить модель науки о данных, чтобы ответить на жизненно важные социальные, финансовые и медицинские вопросы. Во-вторых, такие проекты помогают улучшить портфолио для вашей карьеры. Если вы сможете построить успешную модель анализа данных, которая может предложить практические выводы, вы можете продемонстрировать эти модели в Интернете, создав сайты-портфолио. Работодатели предпочитают проекты, а не целевые заявления.

Бесплатные наборы данных для проектов машинного обучения

Как и специалист по науке о данных, профессионал в области ML также должен работать над самостоятельными проектами, чтобы проверить свои навыки. Если проект окажется успешным, он также станет идеальным компонентом для вашего онлайн или офлайн портфолио ML-проектов. Таким образом, теперь вы понимаете, что развитие науки о данных и ML зависит от структурированных наборов данных. Если бы такие наборы данных были слишком коммерциализированы, исследования и разработки в области науки о данных стали бы полностью ориентированными на корпорации. Для того чтобы исследования в области науки о данных ML были открыты для всех, следующие агентства, учреждения и **платформы предлагают бесплатные наборы данных**:

Data.gov



На сайте Data.gov вы найдете все открытые данные, собранные и обработанные правительством США. Платформа также предлагает ресурсы и инструменты для проведения исследований, создания визуализаций данных, разработки мобильных/веб-приложений и т. д. Среди его наборов данных можно выделить данные об устойчивом землепользовании, данные о сельском жилье, электронные навигационные карты внутренних районов и т.д.

Открытые наборы данных: Kaggle

Kaggle предлагает океан открытых данных и компьютерных кодов для проектов в области науки о данных. Вы можете выбрать Datasets для исходных данных и Code для кодов программирования. Среди трендовых наборов данных на Kaggle – данные AMEX, зрительская аудитория “Симпсонов”, данные для обучения чатботов и т. д.

Сегментные наборы данных: YouTube 8-M

Наборы данных сегментов YouTube 8-M предлагают вам аннотации сегментов, проверенные человеческими аудиторами. На том же портале вы можете получить доступ к набору данных YouTube-8M. Набор содержит 6,1 млн идентификаторов видео, 350 000 часов видео, 2,6 млрд аудио/визуальных признаков, 3863 класса видео и в среднем 3,0 метки на видео.

Реестр открытых данных на AWS

ROD on AWS помогает ученым, изучающим данные, обмениваться и открывать наборы данных, размещенные на ресурсах AWS. Среди интересных наборов данных, которые вы можете найти здесь, – The Cancer Genome Atlas, Foldingathome COVID-19 Datasets, Common Crawl и др.

Репозиторий машинного обучения: UCI

Репозиторий UCI Machine Learning Repository в настоящее время содержит 622

набора данных, которые могут использоваться учеными, занимающимися изучением данных, и инженерами ML для обучения своих моделей искусственного интеллекта. Кроме того, имеется поисковый интерфейс для изучения баз данных. Популярными являются наборы данных Accelerometer dataset, Synchronous Machine dataset, Wikipedia Math Essentials, Turkish Headlines dataset и другие.

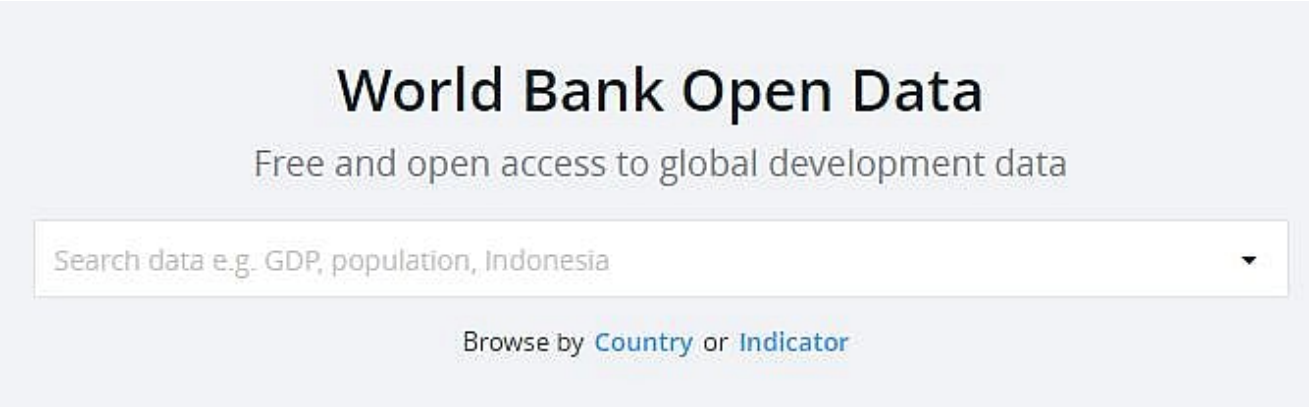
Публичные наборы данных BigQuery: Облако Google

Многие публичные наборы данных хранятся в BigQuery. Google предоставляет к ним бесплатный доступ в рамках программы Google Cloud Public Dataset Program. Однако лимит бесплатных запросов составляет 1 ТБ в месяц. Вы можете выполнять стандартные запросы SQL и унаследованные запросы SQL.

Замечательные публичные наборы данных: GitHub

Awesome Public Datasets – это набор данных с открытым исходным кодом, содержащий ориентированные на конкретную тему публичные данные. Собранные и отсортированные из различных блогов, ответов и отзывов пользователей, они объединяют бесплатные и платные наборы данных по физике, спорту, программному обеспечению, естественному языку и машинному обучению.

Данные Всемирного банка



Открытые данные Всемирного банка – это платформа, на которой вы получаете бесплатный доступ к данным о глобальном развитии. Она также предлагает другие ценные ресурсы, такие как предварительно отформатированные таблицы и отчеты. Вы можете легко найти нужный набор данных по стране или показателю.

FiveThirtyEight

FiveThirtyEight – это американский сайт, который занимается анализом опросов общественного мнения, политикой, экономикой и спортом. Вы можете получить доступ к этим опросам и прогнозам через наборы данных с его платформы. Вы можете скачать наборы данных одним щелчком мыши.

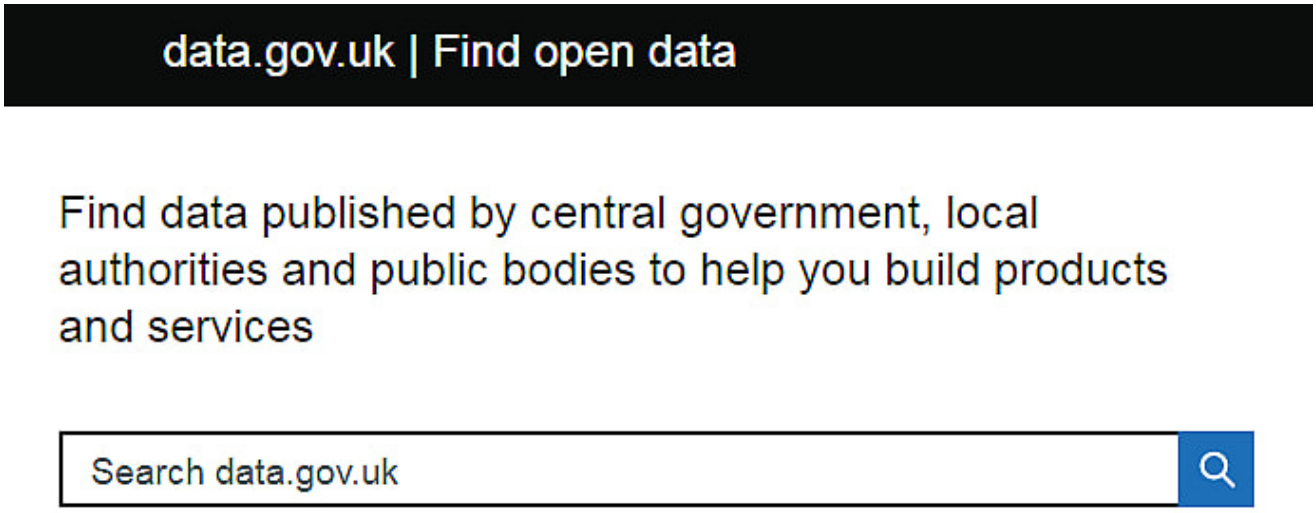
ImageNet

ImageNet – это база данных изображений, из которой исследователи по всему миру могут получить наборы данных с открытым исходным кодом для своих некоммерческих проектов. Здесь изображения организованы на основе иерархии WordNet. Проект играет важную роль в исследованиях глубокого обучения на продвинутом уровне.

Архивы данных: ДАННЫЕ ЮНИСЕФ

С помощью архива данных вы можете получить наборы данных, собранные ЮНИСЕФ по всему миру. Здесь можно найти данные о миграции, перемещении, питании, связи, образовании, здоровье, обучении, смертности, насилии, развитии детей, детских браках, детском труде и различные статистические данные.

Найти открытые данные: Правительство Великобритании



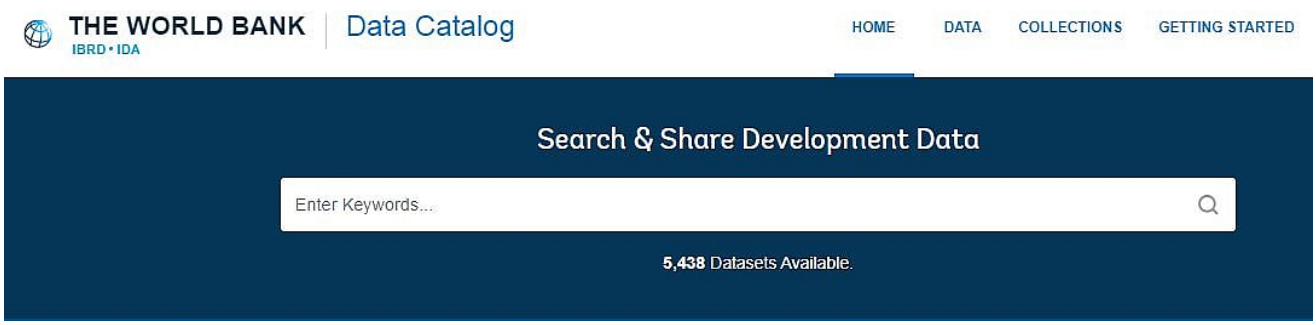
Если вашему проекту нужны данные, опубликованные местными органами власти и центральным правительством Великобритании, то Find Open Data – это портал, который вам стоит посетить. Он охватывает государственные расходы, бизнес, здравоохранение, образование, оборону и другие наборы данных.

Данные и статистика: CDC

Федеральное агентство США Centers for Disease Control and Prevention также предоставляет общественности бесплатные наборы данных для доступа к данным

и статистике с этого портала. Тематика наборов данных включает в себя здоровье окружающей среды, хронические заболевания, рождение и рождаемость, смерть и смертность, продолжительность жизни, травмы и насилие, репродуктивное здоровье, национальные регистрируемые заболевания и т. д.

Каталог данных Всемирного банка

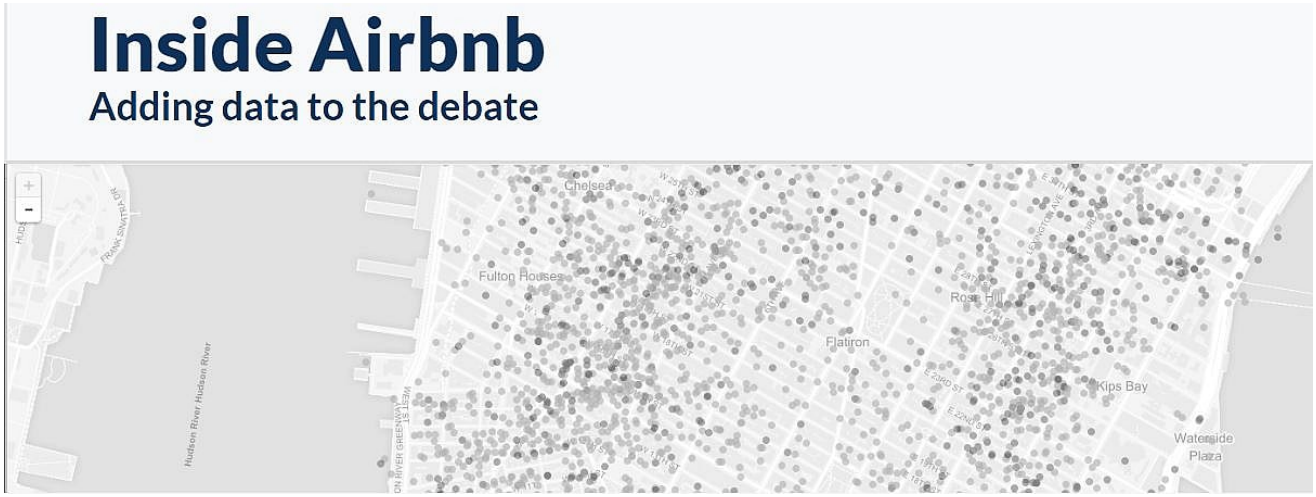


В каталоге данных собраны бесплатные наборы данных, которые делают данные Всемирного банка, связанные с развитием, легкодоступными. Использовать его в различных проектах – проще простого, ведь вы можете без труда найти и загрузить нужную вам информацию. Он содержит более 5000 наборов данных, охватывающих микроданные, финансы и энергетические платформы Всемирного банка.

Данные космической науки НАСА

NASA предоставляет доступ к своим архивным данным на сайте Space Science Data Coordinated Archive. Эта платформа очень полезна для широкой общественности, особенно для людей, работающих в сфере образования и космических исследований. Здесь хранится 400 ТБ цифровых данных, содержащих информацию о 550 космических науках.

Получите данные: Внутри Airbnb



Airbnb – всемирно известная онлайн-площадка для аренды жилья и отдыха. Компания также предлагает сбор данных по различным городам мира от Get the Data. Вы можете просмотреть город, чтобы быстро получить данные. Кроме того, на этом портале можно запросить необходимые данные и ознакомиться с предположениями о них.

Web Data: Amazon Reviews

Тем, кто интересуется маркетинговыми исследованиями и обзорами товаров, стоит воспользоваться набором данных, предоставленным Snap Web Data. Он содержит более 34 миллионов пользовательских отзывов на Amazon с июня 1995 года по март 2013 года. Набор данных содержит обычный текст, информацию о продукте, имя пользователя, оценки и отзыв.

Данные МВФ

На портале данных МВФ можно найти все виды экономических и финансовых данных. Если вы ищете финансовые данные МВФ, статистику по внешнему сектору, основные публикации или данные по микроэкономике, вы можете найти их именно здесь. Кроме того, вы можете использовать фильтр для получения данных по странам.

Данные по рынкам: The Financial Times

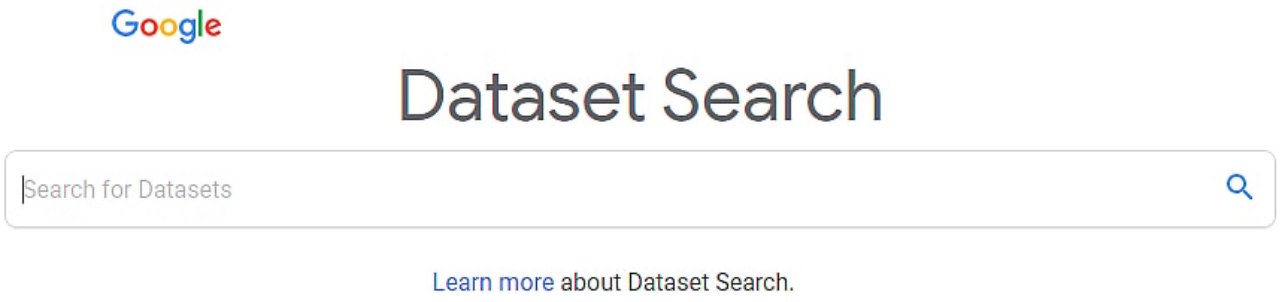
Если вы хотите получить надежные и точные данные о глобальных и региональных рынках акций, Markets Data от The Financial Times поможет вам в этом. С его помощью вы сможете работать с рыночными данными из Америки, Азиатско-

Тихоокеанского региона, Европы, Африки и всего мирового рынка.

Earthdata: NASA

NASA предоставляет полный и открытый доступ к своим научным данным в рамках программы Earth Data, которая поможет вам понять нашу родную планету и реализовать проекты с ее использованием. Вы можете найти бесплатные наборы данных по атмосфере, биосфере, криосфере, человеческим измерениям, поверхности суши, океану, твердой Земле, солнечно-земному взаимодействию и земной гидросфере.

Поиск данных: Google

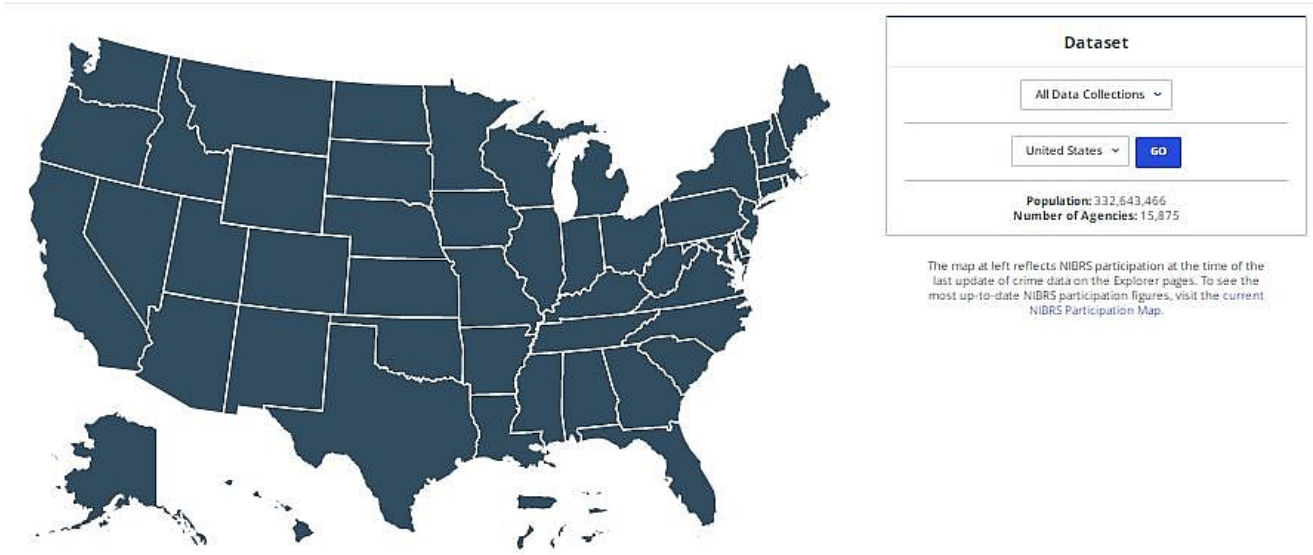


Если вы студент, исследователь или специалист по изучению данных, ищущий наборы данных для поддержки своего проекта, вам поможет портал Dataset Search. Его можно назвать поисковой системой для наборов данных, поскольку он позволяет находить наборы данных, размещенные в различных отчетах по всему Интернету, с помощью поиска по ключевым словам.

Открытые данные: ЦЕРН

Европейская исследовательская организация CERN имеет портал открытых данных, с помощью которого вы можете получить доступ к данным, полученным в результате исследований в CERN. Этот портал содержит два петабайта данных, связанных с физикой частиц. Кроме того, к нему прилагаются приложения и документация, необходимые для анализа данных.

Исследователь данных о преступности: ФБР



Crime Data Explorer (CDE) – это набор данных с открытым исходным кодом от ФБР, цель которого – обеспечить более легкий доступ к обмену криминальными, некриминальными и правоохранительными данными. Кроме того, что эта платформа позволяет находить нужные данные с помощью визуализации и фильтрации по категориям, она также позволяет загружать данные в формате CSV.

Заключительные слова

К настоящему моменту вы ознакомились с поистине исчерпывающим списком высококачественных наборов данных. В статье представлены данные из различных ниш, таких как физические науки, медицинские записи, космические исследования, криминальные записи, рейтинги товаров и т. д. В зависимости от того, какой проект в области науки о данных или машинного обучения вы затеяли, вы можете выбирать. Почти все наборы данных также снабжены соответствующими инструкциями, которые помогут вам в работе над проектом.

Дата Создания
22.02.2024