Источники открытых датасетов для ИИ

Крупные интеграторы данных

Google Dataset Search. Поиск можно начинать отсюда. Все наборы данных отсортированы по:

- актуальности;
- формату файла;
- типу лицензии;
- теме;
- последнему обновлению.
  Базы данных загружаются различными международными организациями, такими как Всемирная Организация Здравоохранения, Statista и Гарвард..

Kaggle — площадке для соревнований по машинному обучению. качество данных может различаться, однако все они совершенно бесплатны. Также есть возможность загрузить в библиотеку свою собственную базу данных. Список датасетов

Data World Каталог, о котором редко упоминают. По способу поиска он похож на поисковик Google. Разница в том, что глубина поиска больше, например, он включает в себя подфайлы, которые могут содержать нужные данные. Это особенно важно при поиске вторичных данных.

UCI Machine Learning Repository Еще один репозиторий с сотнями наборов данных предлагает Калифорнийский университет. Данные в UCI классифицируется по типу задач машинного обучения. Можно найти данные для одномерных и многомерных временных рядов, классификации, регрессии или рекомендательных систем. Некоторые наборы данных в UCI уже очищены и готовы к использованию. Вы можете сортировать пакеты данных по:

- стандартным задачам;
- типам данных;
- области применения;
- предмету.

CMU library Библиотеки CMU Университет Карнеги-Меллона располагает собственной общедоступных наборов данных, которые можно использовать для исследований. Там вы найдете подробные базы данных об американской культуре, музыке и истории, которые не предоставляют другие агрегаторы.

Открытые базы данных на Github

Большая коллекция наборов данных с открытым исходным кодом, разделенных по отраслям. Некоторые из библиотек, которые вы можете там найти

- Congress legislators: база данных людей, избранных в конгресс США Члены Конгресса США (1789 – настоящее время), комитеты Конгресса (1973 – настоящее время), состав комитетов (только текущий), а также президенты и вице-президенты США в формате YAML, JSON и CSV.
- Covid data Полный набор данных COVID-19 – коллекция данных о коронавирусе, которую ведет компания Our World in Data. Ресурс обновляется ежедневно в течение всего периода пандемии COVID-19.

Открытые наборы данных Microsoft Azure

Открытые наборы данных Azure регулярно обновляются и доступны для разработчиков приложений и исследователей. Они содержат правительственные данные США, другие статистические и научные данные, а также информацию из онлайн-сервисов, которую Microsoft собирает о своих пользователях.

- Russian Open Speech To Text

  Коллекция образцов речи, полученных из различных аудиоисточников. Набор данных содержит короткие аудиоклипы на русском языке. Все файлы были преобразованы в opus, за исключением тех, которые служат для проверки. Основная цель набора данных – обучение моделей преобразования речи в текст.

  Russian speech to text (STT) включает:

  1. ~16 миллионов высказываний
  2. ~20 000 часов
  3. 2,3 ТБ (без сжатия в формате .wav в int16), 356 ГБ
- TartanAir

  Одновременная локализация и картирование (SLAM) – одна из самых фундаментальных возможностей, необходимых для роботов. Благодаря повсеместной доступности изображений, визуальная SLAM (V-SLAM) стала важным компонентом многих автономных систем.

  Этот набор данных использует преимущества развивающихся технологий компьютерной графики и направлен на охват различных сценариев со сложными характеристиками при моделировании роботов.

Источники данных по отраслям/задачам

Государственные датасеты:

- Data.gov. Тут находится информация от различных организаций США. Данные могут быть абсолютно разными, от государственного бюджета до отметок в школьном табеле.
- Food Environment Atlas. Включает в себя сведения влиянии многообразия факторов на критерии выбора питания в США и его качества. Из показателей следует отметить расстояние до магазина или ресторана, стоимость продуктов, производителя и другие.
- Chronic disease data. Этот датасет содержит сведения о хронических заболеваниях в США.
- The US National Center for Education Statistics.. Содержит данные об образовательных заведениях и демографии не только в США, но и по всей планете.
- The UK Data Service. Наиболее крупное хранилище информации социальной, экономической и демографической направленности в Великобритании.
- Data USA. Подробная визуализация данных общего доступа в США.
- Данные министерства здравоохранения РФ Данные с этого сайта можно использовать без заключения договора с Министерством здравоохранения РФ. Данные находится в открытом доступе. Информацию можно копировать, публиковать, распространять, видоизменять и объединять с другой информацией, использовать в некоммерческих и коммерческих целя
- Данные министерства культуры РФ Этот ресурс предоставляет информацию о данных на тех же условиях, что и Министерство здравоохранения РФ.

Данные о жилье:

- Boston Housing Dataset. Здесь можно увидеть сведения о жилом фонде в Бостоне, которые собрало бюро, осуществляющее перепись населения США.
- Lincolnshire (UK) House Prices Среднемесячные цены на жилье (£) для графства Линкольншир (Англия, UK) и округов. Все цифры включают сделки с недвижимостью от 10 000 фунтов стерлингов до более чем 1 млн. Данные отфильтрованы по типам домов. Набор данных обновляется ежемесячно за 12-месячный период.
- Zillow Housing data Этот набор данных состоит из нескольких датасетов:

    - Цена дома (Home values) – скорректированный на сезон показатель стоимости дома и изменения на рынке жилья в данном регионе.

  В этом датасете используется мерка ZHVI – Zillow Home Values Index.

  Существует ZHVI верхнего уровня (стоимость домов в диапазоне от 65-го до 95-го процентиля для данного региона) и ZHVI нижнего уровня (стоимость домов в

диапазоне от 5-го до 35-го процентиля для данного региона). Zillow также публикует ZHVI для всех типов домов и апартаментов, учитывая стоимость, количество спален и метраж.

- Прогноз цены дома (Zillow Home Values Forecast) – прогноз индекса стоимости жилья Zillow (ZHVI) на один год. ZHVF создается с использованием среза данных ZHVI по всем домам и доступен как в необработанном, так и в скорректированном виде.
- Аренда (Rentals) – показатель рыночной ставки арендной платы в данном регионе. ZORI (Zillow Observed Rent Index) – индекс арендной платы, который определяется по всей выборке арендного жилья, обеспечивая репрезентативность данных для всего рынка аренды.

Индекс рассчитывается в долларах путем вычисления среднего значения объявленной арендной платы, которая попадает в диапазон от 40-го до 60-го процентиля для всех домов и квартир в данном регионе. Подробную информацию можно найти в методологии ZORI.

Экономика и финансы:

- Quandl. Является неплохим источником информации экономической и финансовой направленности. Используется для строительства прогнозных моделей различных данных экономики или котировок акций.
- **Word Bank Open Data**. Включает определенные информационные комплексы, в которых отражается демографическая ситуация, разнообразные экономические показатели и индикаторы развития по всему миру.
- **IMF Data**. Содержит сведения международного валютного фонда о мировых финансах, долговых критериях, резервах валют, инвестиционные рекомендации и стоимость основных сырьевых товарах.
- Financial Times Market Data. Наиболее точная информация о финансовом рынке по всему миру, в том числе индексы стоимости акций, товаров и валют.
- **Google Trends**. Здесь можно узнать и проанализировать сведения по активности поисковых систем в сети.
- American Economic Association (AEA). Неплохое место для поиска информации о макроэкономических показателях США.

Компьютерное зрение:

- **xView**. Является самым крупным из всех наборов воздушных снимков земли общего доступа. Здесь содержатся картинки разных сцен со всех уголков нашей планеты, которые аннотированы при помощи различных ограничений.
- **Labelme**. Включает большое количество аннотированных картинок.

- ImageNet. Датасет, где можно найти изображения для вновь созданных алгоритмов.
- LSUN. Массив картинок, отсортированных по различным критериям.
- MS COCO. Здесь можно найти все, что потребуется для обнаружения и сегментации объектов.
- Visual Genome.. Размеры датасета с подробно аннотированными изображениями являются самыми крупными.
- Google’s Open Images. Включает коллекцию из более чем 9 миллионов URL-адресов, имеющих метки и охватывающих большое количество категорий под лицензией Creative Commons.
- **Labelled Faces in the Wild**. Включает изображения более 10000 человеческих лиц для применения приложений, в основе которых лежит распознавание лиц.
- Stanford Dogs Dataset. Анализ датасета позволит распознать изображения из определенных пород собак.
- Indoor Scene Recognition. Один из наиболее больших датасетов в плане узнавания интерьеров. В нем содержится 67 категорий включающих 15 620 картинок.
- CIFAR-10 Набор данных CIFAR-10 состоит из 60 000 цветных изображений 32x32 в 10 классах, по 6000 изображений в каждом классе. Он содержит 50 000 обучающих и 10 000 тестовых изображений. Изображения разделены на пять обучающих и одну тестовую партию по 10 000 изображений. Тестовая партия включает в себя 1000 случайно выбранных изображений из каждого класса. Обучающие партии содержат остальные изображения в случайном порядке. Однако, некоторые из обучающих партий могут содержать больше изображений из одного класса, чем из другого. Между собой обучающие партии включают 5000 изображений из каждого класса.
- CityScapes Это новый масштабный набор данных, который содержит разнообразные стерео видеопоследовательности, записанные в уличных сценах из 50 городов. В них содержатся высококачественные аннотации на уровне пикселей (pixel-level) для 5000 кадров, в дополнение к набору из 20 000 слабо аннотированных кадров.
- Objectron Набор данных Objectron представляет собой коллекцию коротких, ориентированных на объект видеоклипов, которые сопровождаются метаданными AR-сессии. Они включают в себя расположения камеры, разреженные облака точек и характеристику плоских поверхностей в окружающей среде. В каждом видеоролике камера перемещается вокруг объекта, снимая его под разными углами. Данные содержат аннотированные вручную трехмерные ограничительные рамки для каждого объекта, которые описывают его положение, ориентацию и размеры. Набор данных состоит из 15 000 аннотированных видеоклипов, дополненных более чем 4 млн аннотированных изображений в следующих категориях: велосипеды, книги, бутылки, камеры, коробки с крупами, стулья, чашки, ноутбуки и обувь. Для обеспечения географического разнообразия набор

данных собран в 10 странах на 5 континентах. Вместе с «датой» ресурс предлагает решение для обнаружения 3D-объектов четырех категорий: обуви, стульев, кружек и камер. Модели, приведенные в качестве примера, обучены с использованием данных Objectron и выпущены в MediaPipe.

- VisualData. Датасеты для компьютерного зрения, разбитые по категориям. Доступен поиск.
- COIL100. 100 разных объектов, изображённых под каждым углом в круговом обороте.
- База данных MNIST представляет собой набор моделей для распознавания рукописных цифр. Он содержит обучающий набор из более чем 60 000 примеров и тестовый—из 10 000. На веб-сайте вы также найдете таблицу, в которой сравнивается эффективность различных типов классификаторов, применяемых к этому набору данных. Даже новичок может использовать MNIST для обучения своей модели глубокого обучения.
- Кинетика-700 Большой высококачественный набор видео, содержащий URL-ссылки примерно на 650000 видеоклипов Youtube, которые охватывают 700 классов действий человека. Видео включают взаимодействия человека с объектом, а также взаимодействия человека с человеком. Набор данных Kinetics отлично подходит для обучения модели распознавания действий человека.

Анализ тональности текста:

- Multidomain sentiment analysis dataset.. Достаточно возрастной проект, в котором содержится информация о товарах, купленных на Amazon.
- IMDB reviews. Староватый, относительной небольшой датасет для бинарного анализа тональности.
- Stanford Sentiment Treebank. Проект Стенфортского университета, где анализируют тональность.
- Sentiment140. Модный портал, в котором можно найти множество твитов с удалёнными смайликами.
- Twitter US Airline Sentiment. Здесь находятся данные из Twitter обо всех компаниях авиаперевозчиках США.
- Sentiment analysis Набор различных датасетов, каждый из которых содержит необходимую информацию для анализа тональности текста. Так, данные, взятые с IMDb – это бинарный набор для анализа настроений. Он состоит из 50 000 отзывов из базы данных фильмов (IMDb), помеченных как положительные или отрицательные. Данные содержат только поляризованные отзывы. Отрицательный отзыв имеет оценку ≤ 4 из 10, положительный – ≥ 7 из 10. На каждый фильм включается не более 30 рецензий. Модели оцениваются по точности.
- SMS спам Коллекция SMS-спама v.1 – общедоступный набор SMS-сообщений с метками, которые были собраны для исследования спама с мобильных

телефонов. Данные состоят из 5574 англоязычных, реальных и неконсолидированных сообщений, помеченных как легитимные (ham) или спам. Сообщения SMS-спама были вручную извлечены с веб-сайта Grumbletext. Это британский форум, на котором пользователи мобильных телефонов публично заявляют о спамовых SMS-сообщениях. Идентификация текста спам-сообщений в претензиях – сложная и трудоемкая задача. Она включает тщательное сканирование сотен веб-страниц.

- WikiQA WikiQA представляет собой набор пар вопросов и предложений. Они были собраны и аннотированы для исследования ответов на вопросы в открытых доменах. Большинство предыдущих работ по выбору предложений для ответа сосредоточено на наборе данных, созданном на основе данных TREC-QA, который включает вопросы, созданные редакторами, и предложения-кандидаты для ответа, отобранные по совпадению содержательных слов в вопросе. WikiQA создана с использованием более естественного процесса. Она включает вопросы, для которых не существует правильных предложений, что позволяет исследователям работать над триггером ответа, критически важным компонентом любой системы QA.

Обработка естественного языка:

- **HotspotQA Dataset**. Ресурс, в котором содержатся вопросы и ответы. С его помощью можно создать систему стандартных ответов.
- Google Books Ngrams. Включает коллекцию слов из книги Google.
- Wikipedia Links data. Этот проект построен из веб-страниц, причем на каждой имеется одна ссылка на Википедию и ее якорный текст аналогичен заголовку страницы.
- Gutenberg eBooks List.. Датасет с аннотированным списком электронных книг проекта «Гутенберг».
- Jeopardy. Содержит архивные данные одноименной телевизионной викторины.
- Rotten Tomatoes Reviews. Здесь находятся рецензии в количестве 480 тысяч штук с Rotten Tomatoes.
- UCI’s Spambase Крупный датасет, в котором находятся спам-письма.
- Text classification TREC – это набор данных для классификации вопросов, который состоит из открытых вопросов, основанных на фактах. Они разделены на широкие семантические категории. Датасет имеет шестиклассную (TREC-6) и пятидесятиклассную (TREC-50) версии. Обе версии включают 5452 обучающих и 500 тестовых примеров.
- Amazon Reviews dataset Этот набор данных состоит из нескольких миллионов отзывов покупателей Amazon и их оценок. Датасет используется для возможности обучения fastText, анализируя настроения покупателей. Идея состоит в том, что несмотря на огромный объем данных – это реальная бизнес-задача. Модель

обучается за считанные минуты. Именно это отличает Amazon Reviews от аналогов.

- Yelp dataset Набор данных Yelp – это множество предприятий, отзывов и пользовательских данных, которые можно применить в Pet-проекте и научной работе. Также можно использовать Yelp для обучения студентов во время работы с базами данных, при изучении NLP и в качестве образца производственных данных. Датасет доступен в виде файлов JSON и является «классикой» в обработке естественного языка.
- Enron Dataset. Данные электронной почты от высшего руководства Enron.
- Blogger Corpus. Коллекция из 681 288 постов с Blogger. Каждый блог содержит как минимум 200 вхождений часто используемых английских слов.
- WordNet—это лексическая база данных, содержащая слова всех частей речи, сгруппированных в наборы синонимов. Такая структура делает ее превосходным инструментом для обработки естественного языка и лингвистических исследований.
- 20 Newsgroups—это набор данных, состоящий из более чем 18 000 текстовых документов из 20 различных групп новостей, включая спорт, технологии, искусство, развлечения и т. д.

Автопилоты:

- Berkeley DeepDrive BDD100k.. В настоящий момент является самым большим датасетом для автопилотов. В нем содержится множество видеозаписей вождения, при разнообразных ситуациях.
- Baidu Apolloscapes.. Ресурс с функцией распознавания 26 семантически разных объектов. Это могут быть машины, велосипеды, пешеходы, здания, уличные фонари и т. д.
- **Comma.ai**. Здесь содержится информация об основных параметрах машины, находящейся в движении.
- Oxford’s Robotic Car. Проект включает около 100 повторения одного и того же маршрута, которые были запечатлены за один год в Оксфорде. На маршруте явно прослеживаются разные условия: трафик, погода, пешеходы, ремонт дороги и т.д.
- Cityscape Dataset. Скачав этот датасет, можно найти сто записей с уличных камер из 50 городов.
- KUL Belgium Traffic Sign Dataset Информация, содержащая аннотации к тысячам бельгийских светофоров.
- ONCE Dataset Набор данных ONCE – крупномасштабный набор данных автономного вождения с аннотациями 2D и 3D объектов.

  Включает в себя:

  1. 1 млн кадров LiDAR, 7 млн изображений с камер.
  2. 200 км² регионов вождения, 144 часа вождения.

3. 15 000 полностью аннотированных сцен с 5 классами: автомобиль, автобус, грузовик, пешеход, велосипедист.
4. Разнообразные условия: день/ночь, солнце/дождь, город/пригород.

- Ford AV dataset Ford AV dataset создан в рамках программы AWS Public Dataset Program. Представленные данные организованы на основе временных рядов. Все разделы содержат подразделы для каждого транспортного средства и карт. Каждый подраздел Vehicle включает журналы в формате rosbag, изображения PNG и файлы калибровки всех датчиков. Калибровочные данные для каждого автомобиля предоставляются отдельно.
- Canadian Adverse Driving Conditions Dataset CADC Dataset нацелен на продвижение исследований по улучшению самостоятельного вождения в неблагоприятных погодных условиях. Это первый публичный датасет, который посвящен реальным данным о вождении в снежных погодных условиях.
  Включает в себя:
  1. 56 000 изображений с камер
  2. 7 000 разверток LiDAR
  3. 75 сцен по 50-100 кадров в каждой
  4. 10 классов аннотаций
  5. 28 194 Автомобиля
  6. 62 851 Пешеход
  7. 20 441 Грузовик
  8. 4867 Автобусов
  9. 4808 Мусорных контейнеров
  10. 3205 Объектов, направляющих движение
  11. 705 Велосипедов
  12. 638 Пешеходов с объектом
  13. 75 Лошадей и колясок
  14. 26 Животных
  15. Полный набор датчиков: 1 LiDAR, 8 камер, постобработанный GPS/IMU
  16. Неблагоприятные погодные условия вождения, включая снег
- LISA. Laboratory for Intelligent & Safe Automobiles, UC San Diego Datasets. Датасет с дорожными знаками, светофорами, распознанными средствами передвижения и траекториями движения.
- Bosch Small Traffic Light Dataset. Датасет с 24 000 аннотированных светофоров.
- LaRa Traffic Light Recognition. Ещё один датасет для распознавания светофоров.
- WPI datasets. Датасет для распознавания светофоров, пешеходов и дорожной разметки.

Медицинские данные:

- **MIMIC-III**. Датасет содержащий обезличенную информацию о состоянии здоровья около 40 тысяч больных, которые подвергаются интенсивной терапии. Он включает карту пациента, показатели жизненной активности, принимаемые лекарства, прогноз лечения и т.д.
- Health Science Library Клинические данные являются основным источником большинства медицинских исследований. Они собираются в ходе текущего лечения пациентов или в рамках официальной программы клинических исследований.

  HSL предлагает клинические данные шести основных типов:

  1. Электронные медицинские карты
  2. Административные данные
  3. Данные о претензиях
  4. Регистры пациентов/заболеваний
  5. Медицинские опросы
  6. Данные клинических исследований
- DeepLesion Датасет хронических заболеваний в США (US chronic diseases). DL предлагает набор данных, которые были получены в результате более 10 тыс. исследований на 4 тыс. уникальных пациентов. Данные включают в себя информацию о различных типах поражений, таких как: узелки в легких, опухоли печени, увеличенные лимфатические узлы и т.д. Используя DeepLesion, мы обучаем универсальный детектор поражений, который может находить все их типы поражений с помощью единой унифицированной системы.

Аудио

- VoxCeleb—Обширный набор данных размером 150 МБ, состоящий из почти 2000 часов речи и предназначенный для определения личности говорящего. Он содержит около 100 000 высказываний 1251 знаменитости, взятых из видео на YouTube. Данные почти равномерно распределены по полу (мужчины составляют 55%). Знаменитости различаются по акцентам, профессиям и возрасту. Наборы для разработки и тестирования не содержат совпадений.
- На LibriSpeech вы найдете около 1000 часов устной английской речи частотой 16 кГц, полученной из аудиокниг. Данные взяты из аудиокниг проекта LibriVox, их размер составляет около 60 ГБ
- Free Spoken Digit Dataset состоит из устных цифровых записей на частоте 8 кГц с почти минимальной тишиной в начале и в конце. Набор данных распространяется с открытым исходным кодом. тот открытый набор данных был создан для определения цифр, произносимых в аудиосемплах. На данный момент он содержит: 3 говорящих, 1500 записей (по 50 с каждой цифрой на говорящего), а также вариации английского произношения.

- Common Voice от Mozilla содержит сотни тысяч записей человеческого голоса. Каждый посетитель веб-сайта Common Voice может внести свой вклад, записывав свой собственный голос.
- CREMA-D — датасет для распознавания эмоций по записи голоса.
- Распознавание пола по голосу — эта база данных была создана, чтобы идентифицировать голос как мужской или женский, основываясь на акустических свойствах голоса и речи. Набор данных состоит из 3168 записанных голосовых сэмплов, собранных от мужчин и женщин. (Kaggle)
- YouTube 8M содержит более 6 миллионов видео, метки и около 2,6 миллиарда аудио и визуальных признаков.
- В AudioSet от Google можно найти миллионы помеченных 10-секундных клипов, выбранных из видео YouTube.
- Free Music Archive—это набор данных для анализа музыки. и состоит из полноразмерного HQ-аудио, предварительно вычисленных характеристик, а также метаданных трека и пользовательского уровня. Этот открытый набор данных был создан для оценки нескольких задач поиска музыкальной информации
- Spoken Wikipedia Corpora Созданный волонтерами корпус выровненной разговорной Википедии, включающий сотни статей из английской, немецкой и голландской Википедии. Преимущества этого источника данных сводятся к разнообразному кругу читателей и темам. Все аннотации можно отобразить обратно в исходный html.
- VoxForge Открытый набор данных о речи, созданный для сбора транскрибированной речи на таких языках, как английский, немецкий, итальянский, португальский или испанский.
- Million Song Dataset Открытая коллекция характеристик и метаданных для миллиона треков. Набор не содержит аудио, а только извлеченные характеристики. Аудиосемплы можно получить из таких сервисов, как 7digital, используя код, предоставленный Колумбийским университетом.
- TED-LIUM Набор состоит из 1495 аудиозаписей с выступлений TED Talk и их полных расшифровок, созданных компьютерной лабораторией Университета штата Мэн (LIUM).
- Speech Commands Dataset Набор данных размером 1,4 ГБ включает 65 000 односекундных высказываний из 30 коротких слов, выполненных тысячами разных людей. Выпущен под лицензией Creative Commons-BY 4.0 и разработан для создания простых, но полезных голосовых интерфейсов с общими словами, такими как «да», «нет», цифры и направления движения.
- CHIME Этот набор размером около 4 ГБ предназначен для решения задач по распознаванию речи в шумной обстановке. Он содержит реальные, смоделированные и чистые голосовые записи. Реальные представлены 9000 записями 4 говорящих в 4 шумных местах, смоделированные созданы путем наложения нескольких сред поверх речевых высказываний, а чистые записаны без лишних шумов. Скачать этот набор можно здесь.

- TIMIT Corpus Размер корпуса — 440 МБ. Его данные можно применять для акустико-фонетических исследований, а также для разработки и оценки систем автоматического распознавания речи. TIMIT содержит широкополосные записи 630 носителей восьми основных диалектов американского английского, каждый из которых читает десять предложений с фонетически богатым звучанием. Он включает синхронизированные по времени орфографические, фонетические и словесные транскрипции, а также 16-битный файл речевого сигнала с частотой 16 кГц для каждого высказывания.
- Multimodal EmotionLines Dataset (MELD) MELD — улучшенная и расширенная версия набора данных EmotionLines. Он содержит те же экземпляры диалогов, что и EmotionLines, а также аудио и визуальную модальность наряду с текстом. В нем можно найти более 1400 диалогов и 13 000 высказываний из сериала «Друзья», каждое из которых содержит метку эмоции: гнев, отвращение, печаль, радость, нейтральность, удивление и страх. Скачать этот набор можно здесь.

Наборы данных, включающие звуки окружающей среды

- AudioSet Содержит 632 класса звуковых событий и коллекцию из 2 084 320 помеченных вручную звуковых клипов длиной по 10 секунд, взятых из видео на YouTube. Чтобы скачать этот набор, перейдите по ссылке на GitHub.
- Mivia Audio Events Dataset Включает 6 000 событий, таких как разбивание стекла, выстрелы и крики, разделенных на обучающий набор из 4200 событий и тестовый — из 1800. Чтобы загрузить этот набор данных, нужно зарегистрироваться на сайте Mivia.
- Environmental Audio Datasets Страница включает наборы данных для исследования звуков окружающей среды. Помимо открытых наборов, содержит также частные и коммерческие, а в конце перечислено несколько звуковых онлайн-сервисов, которые можно применять для формирования новых наборов данных для особых исследовательских потребностей.
  Наборы разделены на две таблицы:
  - Таблица звуковых событий содержит наборы данных, подходящие для исследований в области автоматического обнаружения звуковых событий и автоматической маркировки звуков.
  - Таблица акустических сцен включает наборы, которые пригодятся для распознавания контекста на основе звука и классификации акустических сцен.
- FSD & Freesound Иерархическая коллекция из более чем 600 звуковых классов, дополненная 297 159 аудиосемплами от Freesound. В результате этого объединения было создано 678 511 аннотаций кандидатов, которые отражают потенциальное присутствие источников звука в аудио клипах. FSD включает множество повседневных звуков: человеческая речь, звуки животных, музыка и звуки, издаваемые вещами — и все это под лицензией Creative Commons. Набор

данных предназначен для помощи исследованиям, которые позволят машинам слышать и интерпретировать звук подобно людям. Freesound — это платформа для совместного создания аудиоколлекций, помеченных вручную и основанных на контенте Freesound.

- Urban Sound Classification Этот набор данных размером 6 ГБ содержит 8732 помеченных звуковых отрывка из 10 звуковых классов: шум кондиционера, автомобильный гудок, играющие дети, лай собаки, шум бурения и двигателя, выстрел, отбойный молоток, сирена и уличная музыка. Длина каждого — около 4 секунд. Данные содержат такие атрибуты, как ID — уникальный идентификатор звукового отрывка и Class — тип звука.
- Urban Sound Dataset Этот набор включает 1302 звуковых записей, в каждой из которых отмечены начало и конец звукового события из 10 классов: шум кондиционера, автомобильный гудок, играющие дети, лай собаки, шум бурения и двигателя, выстрел, отбойный молоток, сирена и уличная музыка. Некоторые записи содержат несколько звуковых событий, но для каждого файла помечены только события из одного класса. Классы взяты из таксономии городских звуков.
- Bird Audio Detection challenge Набор предназначен для создания надежного и масштабируемого алгоритма обнаружения птиц. Для решения этой задачи используются наборы данных размером 5,4 ГБ, взятые из реальных проектов по мониторингу биоакустики, и объективная стандартизированная структура оценки.

Системы рекомендаций

- Amazon Product Data содержит метаданные и отзывы о миллионах товаров, проданных на Amazon. Это невероятный ресурс для всех, кто интересуется системами рекомендаций.
- MovieLens—это веб-сайт, который предоставляет своим пользователям персонализированные рекомендации по фильмам. У них также есть набор данных с открытым исходным кодом, который вы можете использовать для обучения своего проекта.
- Jester Collaborative Filtering Dataset содержит более 4 миллионов оценок 100 шуток от 73 421 пользователя. Смейтесь до упаду, занимаясь своими исследованиями по МО.

Компания «ПромоДата» мониторит цены на продукты и непродовольственные FMCG-товары по всей России. В бесплатной версии можно получить эксельку с самыми популярными товарами в Москве за последний месяц: сколько они стоят в разных магазинах.

1. Заходите на сайт Promodata.ru.
2. Внизу вводите адрес почты и говорите «Получить пример отчёта».

3. Если нужны более глубокие данные, выбираете нужные штрихкоды и идёте на https://promodata.ru/pokodu.
4. Загружаете файлик со штрихкодами (экселька, csv). Получаете на почту детальный отчёт по этим штрихкодам.