The 50 Best Public Datasets for Machine Learning

STACY STANFORD OCTOBER 02, 2018

What are some open datasets for machine learning? After scrapping the web for hours after hours, we have created a great cheat sheet for high quality and diverse machine learning datasets.

October 2, 2018 by Stacy Stanford

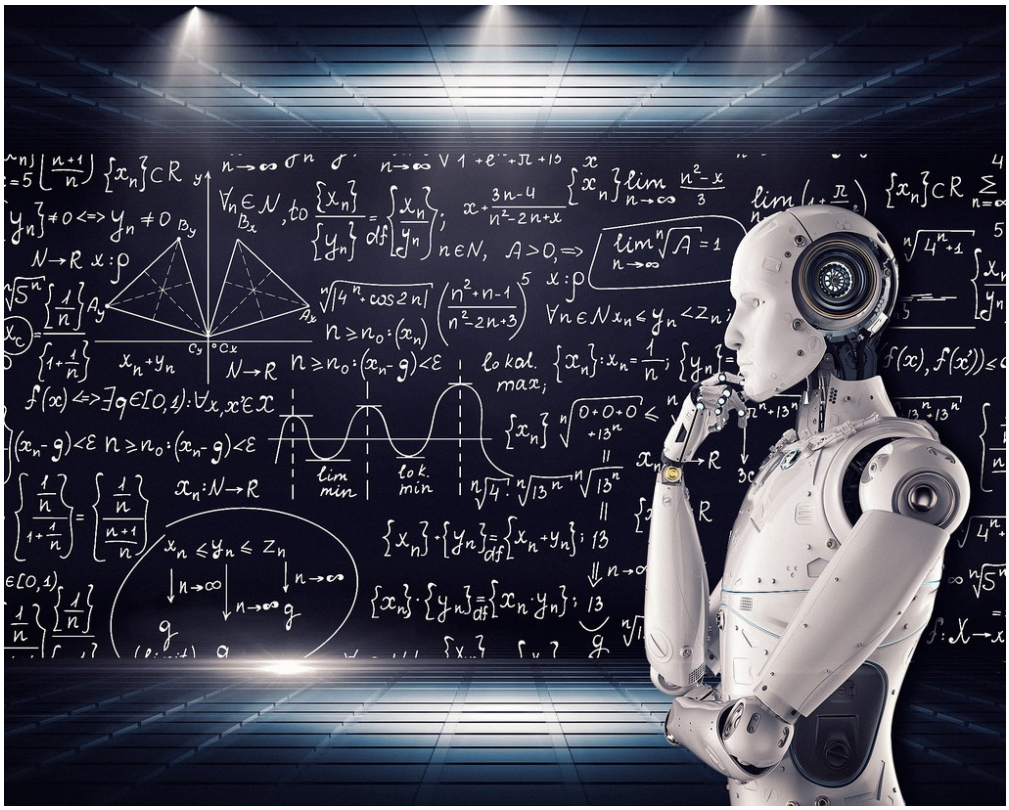

Image Credits: Flickr via www.vpnsrus.com

First, a couple of pointers to keep in mind when searching for datasets. According to Carnegie Mellon University.

> A dataset should not be messy, because you do not want to spend a lot of time cleaning data.

> A dataset should not have too many rows or columns, so it is easy to work with.

> The cleaner the data, the better — cleaning a large data set can be very time consuming.

> There should be an interesting question, which in turn can be answered with data.

Dataset Finders

Kaggle: A data science site that contains a variety of externally contributed interesting datasets. You can find all kinds of niche datasets in its master list, from ramen ratings to basketball data to and even seattle pet licenses.

UCI Machine Learning Repository: One of the oldest sources of datasets on the web, and a great first stop when looking for interesting datasets. Although the data sets are user-contributed, and thus have varying levels of cleanliness, the vast majority are clean. You can download data directly from the UCI Machine Learning repository, without registration.

VisualData: Discover computer vision datasets by category, it allows searchable queries.

General Datasets

Public Government datasets

Data.gov: This site makes it possible to download data from multiple US government agencies. Data can range from government budgets to school

performance scores. Be warned though: much of the data requires additional research.

Food Environment Atlas: Contains data on how local food choices affect diet in the US.

School system finances: A survey of the finances of school systems in the US.

Chronic disease data: Data on chronic disease indicators in areas across the US.

The US National Center for Education Statistics: Data on educational institutions and education demographics from the US and around the world.

The UK Data Service: The UK's largest collection of social, economic and population data.

Data USA: A comprehensive visualization of US public data.

Finance & Economics

Quandl: A good source for economic and financial data — useful for building models to predict economic indicators or stock prices.

World Bank Open Data: Datasets covering population demographics, a huge number of economic, and development indicators from across the world.

IMF Data: The International Monetary Fund publishes data on international finances, debt rates, foreign exchange reserves, commodity

prices and investments.

Financial Times Market Data: Up to date information on financial markets from around the world, including stock price indexes, commodities and foreign exchange.

Google Trends: Examine and analyze data on internet search activity and trending news stories around the world.

American Economic Association (AEA): A good source to find US macroeconomic data.

Machine Learning Datasets:

Images

Labelme: A large dataset of annotated images.

ImageNet: The de-facto image dataset for new algorithms, organized according to the WordNet hierarchy, in which hundreds and thousands of images depict each node of the hierarchy.

LSUN: Scene understanding with many ancillary tasks (room layout estimation, saliency prediction, etc.)

MS COCO: Generic image understanding and captioning.

COIL100 : 100 different objects imaged at every angle in a 360 rotation.

Visual Genome: Very detailed visual knowledge base with captioning of ~100K images.

Google's Open Images: A collection of 9 million URLs to images "that have been annotated with labels spanning over 6,000 categories" under Creative Commons.

Labelled Faces in the Wild: 13,000 labeled images of human faces, for use in developing applications that involve facial recognition.

Stanford Dogs Dataset: Contains 20,580 images and 120 different dog breed categories.

Indoor Scene Recognition: A very specific dataset and very useful, as most scene recognition models are better 'outside'. Contains 67 Indoor categories, and 15620 images.

Sentiment Analysis

Multidomain sentiment analysis dataset: A slightly older dataset that features product reviews from Amazon.

IMDB reviews: An older, relatively small dataset for binary sentiment classification features 25,000 movie reviews.

Stanford Sentiment Treebank: Standard sentiment dataset with sentiment annotations.

Sentiment140: A popular dataset, which uses 160,000 tweets with emoticons pre-removed.

Twitter US Airline Sentiment: Twitter data on US airlines from February 2015, classified as positive, negative, and neutral tweets

Natural Language Processing

HotspotQA Dataset: Question answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems.

Enron Dataset: Email data from the senior management of Enron, organized into folders.

Amazon Reviews: Contains around 35 million reviews from Amazon spanning 18 years. Data include product and user information, ratings, and the plaintext review.

Google Books Ngrams: A collection of words from Google books.

Blogger Corpus: A collection 681,288-blog posts gathered from blogger.com. Each blog contains a minimum of 200 occurrences of commonly used English words.

Wikipedia Links data: The full text of Wikipedia. The dataset contains almost 1.9 billion words from more than 4 million articles. You can search by word, phrase or part of a paragraph itself.

Gutenberg eBooks List: Annotated list of ebooks from Project Gutenberg.

Hansards text chunks of Canadian Parliament: 1.3 million pairs of texts from the records of the 36th Canadian Parliament.

Jeopardy: Archive of more than 200,000 questions from the quiz show Jeopardy.

SMS Spam Collection in English: A dataset that consists of 5,574 English SMS spam messages

Yelp Reviews: An open dataset released by Yelp, contains more than 5 million reviews.

UCI's Spambase: A large spam email dataset, useful for spam filtering.

Self-driving

Berkeley DeepDrive BDD100k: Currently the largest dataset for self-driving AI. Contains over 100,000 videos of over 1,100-hour driving experiences across different times of the day and weather conditions. The annotated images come from New York and San Francisco areas.

Baidu Apolloscapes: Large dataset that defines 26 different semantic items such as cars, bicycles, pedestrians, buildings, streetlights, etc.

Comma.ai: More than 7 hours of highway driving. Details include car's speed, acceleration, steering angle, and GPS coordinates.

Oxford's Robotic Car: Over 100 repetitions of the same route through Oxford, UK, captured over a period of a year. The dataset captures different combinations of weather, traffic and pedestrians, along with long-term changes such as construction and roadworks.

Cityscape Dataset: A large dataset that records urban street scenes in 50 different cities.

CSSAD Dataset: This dataset is useful for perception and navigation of autonomous vehicles. The dataset skews heavily on roads found in the

developed world.

KUL Belgium Traffic Sign Dataset: More than 10000+ traffic sign annotations from thousands of physically distinct traffic signs in the Flanders region in Belgium.

MIT AGE Lab: A sample of the 1,000+ hours of multi-sensor driving datasets collected at AgeLab.

LISA: Laboratory for Intelligent & Safe Automobiles, UC San Diego Datasets: This dataset includes traffic signs, vehicles detection, traffic lights, and trajectory patterns.

Bosch Small Traffic Light Dataset: Dataset for small traffic lights for deep learning.

LaRa Traffic Light Recognition: Another dataset for traffic lights. This is taken in Paris.

WPI datasets: Datasets for traffic lights, pedestrian and lane detection.

Clinical

MIMIC-III: Openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.

Note:

If you are aware of other high-quality, public datasets, which you recommend to people in regards to machine learning, deep learning, etc.

Please feel free to suggest them along with the reasons, why they should be included.

If the reason is strong, We will include them in the list. Also, please let us know your experience with using any of these datasets in the comments section.

> Happy machine learning!

Sources:

LibGuides: Machine Learning and AI: Find Datasets
LibGuides: Machine Learning and AI: Find Datasetsguides.library.cmu.edu

Big Data And AI: 30 Amazing (And Free) Public Data Sources For 2018
*Machine learning, artificial intelligence, blockchains, predictive analytics - all amazing technologies which have…*www.forbes.com

takeitallsource/awesome-autonomous-vehicles
*Curated List of Self-Driving Cars and Autonomous Vehicles Resources - takeitallsource/awesome-autonomous-vehicles*github.com

Fueling the Gold Rush: The Greatest Public Datasets for AI
*It has never been easier to build AI or machine learning-based systems than it is today. The ubiquity of cutting edge…*medium.com

https://www.dataquest.io/blog/free-datasets-for-projects

The Best 25 Datasets for Natural Language Processing | Gengo AI
*Where's the best place to look for free online datasets for NLP? We combed the web to create the ultimate cheat sheet…*gengo.ai

awesomedata/awesome-public-datasets

*A topic-centric list of high-quality open datasets in public domains. New PR ☛ ☛☛ - awesomedata/awesome-public-datasets*github.com

Machine Learning | Carnegie Mellon University

*The Machine Learning Department at Carnegie Mellon University is ranked as #1 in the world for AI and Machine Learning…*www.ml.cmu.edu