25 Open Datasets for Deep Learning Every Data Scientist Must Work With

Pranav Dar — March 29, 2018

Beginner Computer Vision Datasets Deep Learning Image Listicle NLP Research & Techno

Introduction

The key to getting better at deep learning (or most fields in life) is practice. Practice on a variety of problems – from image processing to speech recognition. Each of these problem has it's own unique nuance and approach.

But where can you get this data? A lot of research papers you see these days use proprietary datasets that are usually not released to the general public. This becomes a problem, if you want to learn and apply your newly acquired skills.

If you have faced this problem, we have a solution for you. We have curated a list of openly available datasets for your perusal.

In this article, we have listed a collection of high quality datasets that every deep learning enthusiast should work on to apply and improve their skillset. Working on these datasets will make you a better data scientist and the amount of learning you will have will be invaluable in your career. We have also included papers with state-of-the-art (SOTA) results for you to go through and improve your models.

How to use these datasets?

First things first – these datasets are huge in size! So make sure you have a fast internet connection with no / very high limit on the amount of data you can download.

There are numerous ways how you can use these datasets. You can use them to apply various Deep Learning techniques. You can use them to hone your skills, understand how to identify and structure each problem, think of unique use cases and publish your findings for everyone to see!

The datasets are divided into three categories – Image Processing, Natural Language Processing, and Audio/Speech Processing.

Let's dive into it!

Image Datasets

MNIST

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

set of 60,000 examples and a test set of 10,000 examples. It's a good database f

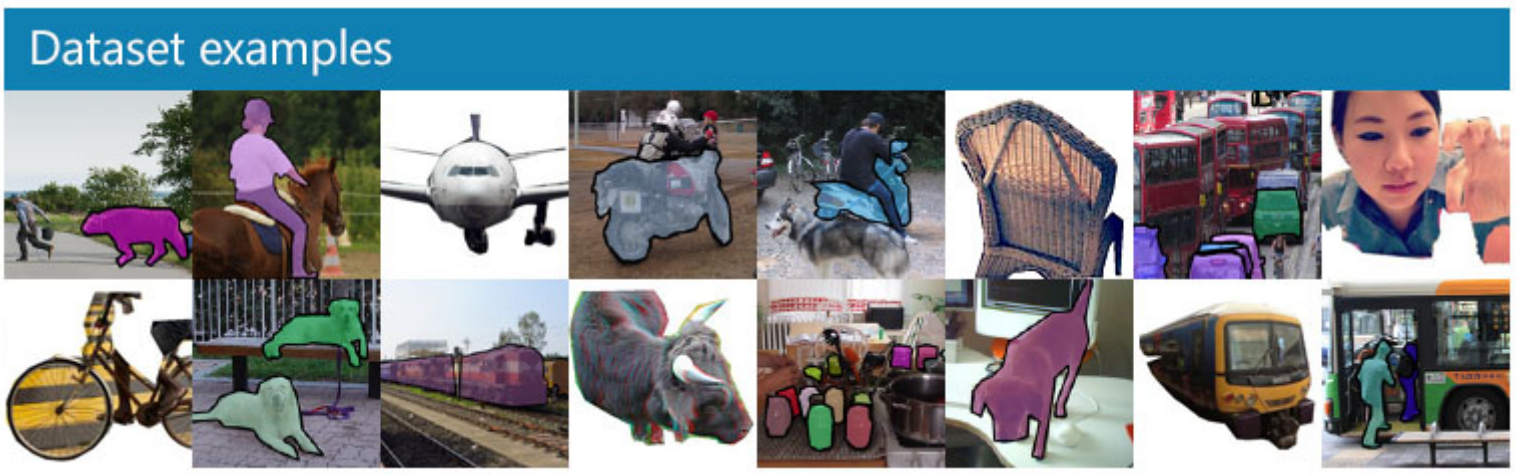
recognition patterns on real-world data while spending minimum time and effo

Size: ~50 MB

Number of Records: 70,000 images in 10 classes

SOTA: Dynamic Routing Between Capsules

MS-COCO



COCO is a large-scale and rich for object detection, segmentation and captioning dataset. It has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Size: ~25 GB (Compressed)

Number of Records: 330K images, 80 object categories, 5 captions per image, 250,000 people with key points

SOTA : Mask R-CNN

Bored with Datasets? Solve real life project on Deep Learning

ImageNet

ImageNet is a dataset of images that are organized according to the WordNet hierarchy. WordNet contains approximately 100,000 phrases and ImageNet has provided around 1000 images on average to illustrate each phrase.

Size: ~150GB

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

Open Images Dataset

Open Images is a dataset of almost 9 million URLs for images. These images have been annotated with image-level labels bounding boxes spanning thousands of classes. The dataset contains a training set of 9,011,219 images, a validation set of 41,260 images and a test set of 125,436 images.

Size: 500 GB (Compressed)

Number of Records: 9,011,219 images with more than 5k labels

SOTA : Resnet 101 image classification model (trained on V2 data): Model checkpoint, Checkpoint readme, Inference code.

VisualQA

VQA is a dataset containing open-ended questions about images. These questions require an understanding of vision and language. Some of the interesting features of this dataset are:

- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Automatic evaluation metric

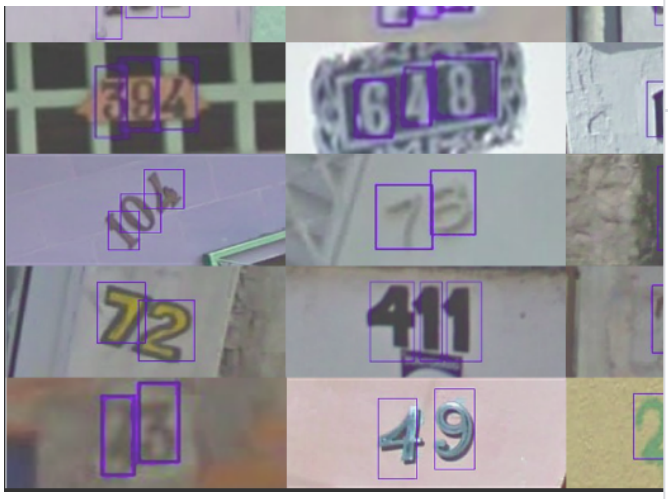
Size: 25 GB (Compressed)

Number of Records: 265,016 images, at least 3 questions per image, 10 ground truth answers per question

SOTA : Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

The Street View House Numbers (SVHN)

25 Open Datasets for Deep Learning Every Data Scientist Must Work With



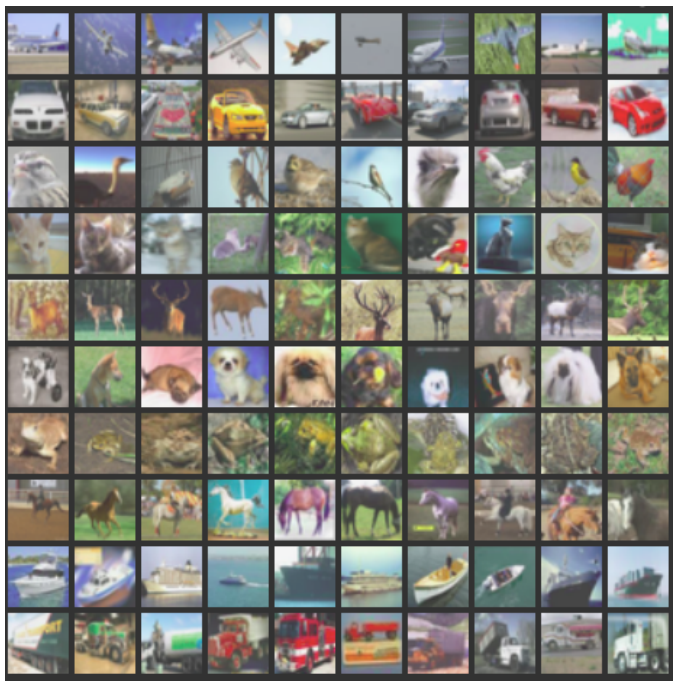
This is a real-world image dataset for developing object detection algorithms. This requires minimum data preprocessing. It is similar to the MNIST dataset mentioned in this list, but has more labelled data (over 600,000 images). The data has been collected from house numbers viewed in Google Street View.

Size: 2.5 GB

Number of Records: 6,30,420 images in 10 classes

SOTA : Distributional Smoothing With Virtual Adversarial Training

CIFAR-10



This dataset is another one for image classification. It consists of 60,000 images of 10 classes (each class is represented as a row in the above image). In total, there are 50,000 training images and 10,000 test images. The dataset is divided into 6 parts – 5 training batches and 1 test batch. Each batch has 10,000 images.

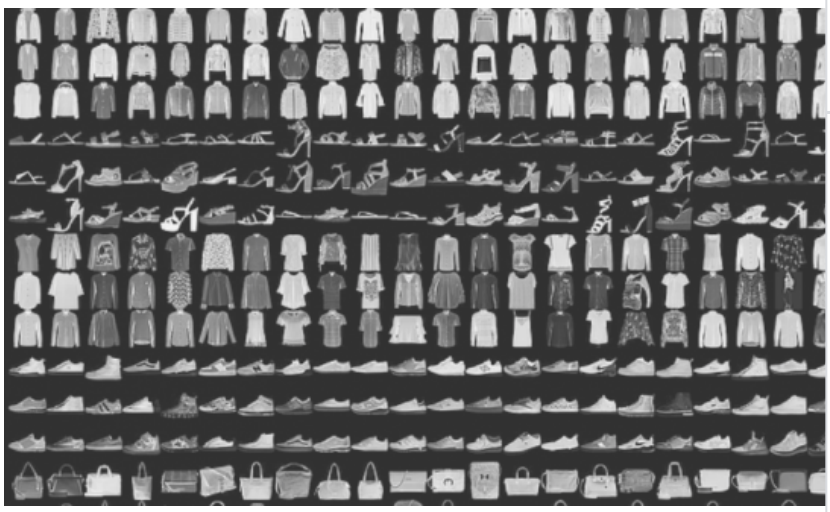
Size: 170 MB

Number of Records: 60,000 images in 10 classes

SOTA : ShakeDrop regularization

Fashion-MNIST

25 Open Datasets for Deep Learning Every Data Scientist Must Work With



Fashion-MNIST consists of 60,000 training images and 10,000 test images. It is a MNIST-like fashion product database. The developers believe MNIST has been overused so they created this as a direct replacement for that dataset. Each image is in greyscale and associated with a label from 10 classes.

Size: 30 MB

Number of Records: 70,000 images in 10 classes

SOTA : Random Erasing Data Augmentation

Natural Language Processing

IMDB Reviews

This is a dream dataset for movie lovers. It is meant for binary sentiment classification and has far more data than any previous datasets in this field. Apart from the training and test review examples, there is further unlabeled data for use as well. Raw text and preprocessed bag of words formats have also been included.

Size: 80 MB

Number of Records: 25,000 highly polar movie reviews for training, and 25,000 for testing

SOTA : Learning Structured Text Representations

Twenty Newsgroups

This dataset, as the name suggests, contains information about newsgroups. To curate this dataset, 1000 Usenet articles were taken from 20 different newsgroups. The articles have typical features like subject lines, signatures, and quotes.

Size: 20 MB

Number of Records: 20,000 messages taken from 20 newsgroups

SOTA : Very Deep Convolutional Networks for Text Classification,

Sentiment140

Sentiment140 is a dataset that can be used for sentiment analysis. A popular dataset, it is perfect to start off your NLP journey.

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

- date of the tweet
- the query
- username of the tweeter
- text of the tweet

Size: 80 MB (Compressed)

Number of Records: 1,60,000 tweets

SOTA : Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentir

WordNet

Mentioned in the ImageNet dataset above, WordNet is a large database of English synsets. Synsets are groups of synonyms that each describe a different concept. WordNet's structure makes it a very useful tool for NLP.

Size: 10 MB

Number of Records: 117,000 synsets is linked to other synsets by means of a small number of "conceptual relations.

SOTA : Wordnets: State of the Art and Perspectives

Yelp Reviews

This is an open dataset released by Yelp for learning purposes. It consists of millions of user reviews, businesses attributes and over 200,000 pictures from multiple metropolitan areas. This is a very commonly used dataset for NLP challenges globally.

Size: 2.66 GB JSON, 2.9 GB SQL and 7.5 GB Photos (all compressed)

Number of Records: 5,200,000 reviews, 174,000 business attributes, 200,000 pictures and 11 metropolitan areas

SOTA : Attentive Convolution

The Wikipedia Corpus

This dataset is a collection of a the full text on Wikipedia. It contains almost 1.9 billion words from more than 4 million articles. What makes this a powerful NLP dataset is that you search by word, phrase or part of a paragraph itself.

Size: 20 MB

Number of Records: 4,400,000 articles containing 1.9 billion words

SOTA : Breaking The Softmax Bottelneck: A High-Rank RNN language Model

The Blog Authorship Corpus

This dataset consists of blog posts collected from thousands of bloggers and has been gathered from blogger.com. Each blog is provided as a separate file. Each blog contains a minimum of 200 occurrences of commonly used English words.

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

SOTA : Character-level and Multi-channel Convolutional Neural Networks for L

Machine Translation of Various Languages

This dataset consists of training data for four European languages. The task her

You can participate in any of the following language pairs:

- English-Chinese and Chinese-English
- English-Czech and Czech-English
- English-Estonian and Estonian-English
- English-Finnish and Finnish-English
- English-German and German-English
- English-Kazakh and Kazakh-English
- English-Russian and Russian-English
- English-Turkish and Turkish-English

Size: ~15 GB

Number of Records: ~30,000,000 sentences and their translations

SOTA : Attention Is All You Need

Engage with real life projects on Natural Language Processing here

Audio/Speech Datasets

Free Spoken Digit Dataset

Another entry in this list for inspired by the MNIST dataset! This one was created to solve the task of identifying spoken digits in audio samples. It's an open dataset so the hope is that it will keep growing as people keep contributing more samples. Currently, it contains the below characteristics:

- 3 speakers
- 1,500 recordings (50 of each digit per speaker)
- English pronunciations

Size: 10 MB

Number of Records: 1,500 audio samples

SOTA : Raw Waveform-based Audio Classification Using Sample-level CNN Architectures

Free Music Archive (FMA)

FMA is a dataset for music analysis. The dataset consists of full-length and HQ audio, pre-computed features, and track and user-level metadata. It an an open dataset created for evaluating several tasks in MIR. Below is the list of csv files the dataset has along with what they include:

- `tracks.csv`: per track metadata such as ID, title, artist, genres, tags and play counts, for all 106,574 tracks.
- `genres.csv`: all 163 genre IDs with their name and parent (used to infer the genre hierarchy and top-level genres).
- `features.csv`: common features extracted with librosa.
- `echonest.csv`: audio features provided by Echonest (now Spotify) for a subset of 13,129 tracks.

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

SOTA : Learning to Recognize Musical Genre from Audio

Ballroom

This dataset contains ballroom dancing audio files. A few characteristic excerpts

format. Below are a few characteristics of the dataset:

- Total number of instances: 698
- Duration: ~30 s
- Total duration: ~20940 s

Size: 14GB (Compressed)

Number of Records: ~700 audio samples

SOTA : A Multi-Model Approach To Beat Tracking Considering Heterogeneous Music Styles

Million Song Dataset

The **Million Song Dataset** is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. Its purposes are:

- To encourage research on algorithms that scale to commercial sizes
- To provide a reference dataset for evaluating research
- As a shortcut alternative to creating a large dataset with APIs (e.g. The Echo Nest's)
- To help new researchers get started in the MIR field

The core of the dataset is the feature analysis and metadata for one million songs. The dataset does not include any audio, only the derived features. The sample audio can be fetched from services like 7digital, using code provided by Columbia University.

Size: 280 GB

Number of Records: PS – its a million songs!

SOTA : Preliminary Study on a Recommender System for the Million Songs Dataset Challenge

LibriSpeech

This dataset is a large-scale corpus of around 1000 hours of English speech. The data has been sourced from audiobooks from the LibriVox project. It has been segmented and aligned properly. If you're looking for a starting point, check out already prepared Acoustic models that are trained on this data set at kaldi-asr.org and language models, suitable for evaluation, at http://www.openslr.org/11/.

Size: ~60 GB

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

VoxCeleb

VoxCeleb is a large-scale speaker identification dataset. It contains around 100,
from YouTube videos. The data is mostly gender balanced (males comprise of 5
accents, professions and age. There is no overlap between the development and
and identifying which superstar the voice belongs to.

Size: 150 MB

Number of Records: 100,000 utterances by 1,251 celebrities

SOTA : VoxCeleb: a large-scale speaker identification dataset

Analytics Vidhya Practice Problems

For your practice, we also provide real life problems and datasets to get your hands dirty. In this section, we've listed down the deep learning practice problems on our DataHack platform.

Twitter Sentiment Analysis

Hate Speech in the form of racism and sexism has become a nuisance on twitter and it is important to segregate these sort of tweets from the rest. In this Practice problem, we provide Twitter data that has both normal and hate tweets. Your task as a Data Scientist is to identify the tweets which are hate tweets and which are not.

Size: 3 MB

Number of Records: 31,962 tweets

Age Detection of Indian Actors

This is a fascinating challenge for any deep learning enthusiast. The dataset contains thousands of images of Indian actors and your task is to identify their age. All the images are manually selected and cropped from the video frames resulting in a high degree of variability interms of scale, pose, expression, illumination, age, resolution, occlusion, and makeup.

Size: 48 MB (Compressed)

Number of Records: 19,906 images in the training set and 6636 in the test set

SOTA: Hands on with Deep Learning – Solution for Age Detection Practice Problem

Urban Sound Classification

This dataset consists of more than 8000 sound excerpts of urban sounds from 10 classes. This practice problem is meant to introduce you to audio processing in the usual classification scenario

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

If you are aware of other open datasets, which you recommend to people starti
datasets, please feel free to suggest them along with the reasons, why they shou

If the reason is good, I'll include them in the list. Let us know your experience wi
section. And happy deep learning!

Not sure, how to start your Deep Learning Journey? Use our Learning Path for D

Learn, engage , compete and get hired!

Related

Top 15 Open-Source Datasets of 2020 that every Data Scientist Should add to their Portfolio!

24 Ultimate Data Science (Machine Learning) Projects To Boost Your Knowledge and Skills (& can be accessed freely)

25+ websites to find datasets for data science projects

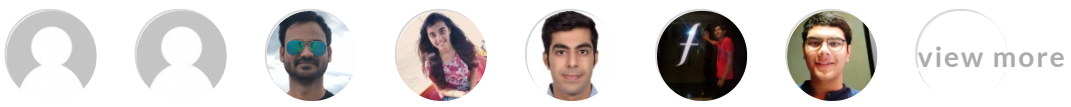
data science projects datasets deep learning image datasets MNIST data Practice Problems text datasets voice datasets

About the Author

Pranav Dar

Senior Editor at Analytics Vidhya. Data visualization practitioner who loves reading and delving deeper into the data science and machine learning arts. Always looking for new ways to improve processes using ML and AI.

Our Top Authors



25 Open Datasets for Deep Learning Every Data Scientist Must Work With

Automated Machine Learning Tool in Action

29 thoughts on "25 Open Datasets for Deep Learning Every

Sujatha Sivaraman says:
March 29, 2018 at 11:48 am

Thanks Pranav. This is indeed going to be useful to strengthen your practice

Reply

Sunny Toms says:
March 29, 2018 at 2:48 pm

Good work Pranav. This is good info for deep learning self learners. Thank you.

Reply

Fredrik says:
March 29, 2018 at 2:53 pm

Very good, thanks!

Reply

Mithlesh Patel says:
March 29, 2018 at 3:50 pm

Thankyou Pranav, very useful for self learners.

Reply

Raghu S says:
March 29, 2018 at 3:57 pm

Sir, where can I get medical image dataset. Please provide necessary information.

Reply

N says:
March 29, 2018 at 4:36 pm

Great job Pranav, very usefuil

Reply

Manish says:
March 29, 2018 at 5:55 pm

Great job Pranav. Keep it.

Reply

Swapna says:
March 29, 2018 at 6:24 pm

Thanks Pranav. Good info for self learners

Reply

Mathias Müller says:
March 30, 2018 at 2:40 am

A machine translation researcher here, Regarding the machine translation data set you present: You clearly have not researched this well. You link randomly to the WMT machine translation shared task of 2011, although there are 7 newer editions, including 2018 (http://statmt.org/wmt18/index.html). Also, the "Attention is All you Need" paper from Google is _not_ state of the art on the 2011 data you are describing, the paper only has results from the 2014 shared task), and only from English to German, and English to French. By the way, the data sets are of course not called "Machine Translation of European Languages". WMT data is comprised of several individual data sets, such as Europarl, JRC Aquis, News Commentary or OpenSubtitles. Please do your homework.

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

that, the below GitHub link contains some datasets that you might find usefu

Reply

Peter says:
March 31, 2018 at 11:55 pm

Thank You!

Reply

ahmed awaad says:
April 01, 2018 at 4:33 am

we need dataset about link quality

Reply

David says:
April 02, 2018 at 1:16 am

We maintain this huge list of color-names: https://github.com/meodai/color-names/ Could also be useful to categorise or name colors

Reply

V V CHAKRADHAR says:
April 02, 2018 at 9:31 pm

Hey Pranav , I am working on Resume Ranking . Can you help me with a good link !!

Reply

Pranav Dar says:
April 03, 2018 at 11:41 am

Hi Ahmed, Can you elaborate a little bit on what you mean by "link quality"?

Reply

Pranav Dar says:
April 03, 2018 at 11:42 am

Hi Mathias, Thanks for taking the time to read through the article and giving us your suggestions! We have updated the link for the machine translation set accordingly. Since the latest version has incorporated a few other languages as well, we have updated the name to reflect that.

Reply

Pranav Dar says:
April 03, 2018 at 12:26 pm

Hi VV, Are you looking for a dataset or a general benchmark? You can check out this article to see how a job site is using ML to find deserving and qualified candidates: https://www.analyticsvidhya.com/blog/2018/02/job-site-machine-learning-uncommon/

Reply

Pranav Dar says:
April 03, 2018 at 12:29 pm

Looks pretty intriguing. Thanks for sharing, David!

Reply

Mathias Müller says:
April 03, 2018 at 2:58 pm

Hi Pranav Thanks for your reply. In retrospect, I do not like the condescending tone of my comment, my apologies. I am mainly taking issue because of the strong title ("Data every data scientist MUST work with"). Let me try to explain things a bit more. The theme of your post is to present individual data sets, say, the MNIST digits. But for machine translation, people usually aggregate and blend different individual data sets. Your section about machine translation is misleading in that it suggests there is a self-contained data set called "Machine Translation of Various Languages". An actual data set for machine translation you could mention is OpenSubtitles: http://opus.nlpl.eu/OpenSubtitles.php. Or Europarl: http://www.statmt.org/europarl/. Or the English

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

April 04, 2018 at 7:41 am

The Wiki Corpus you link to is pay; $245 for the complete dump.

Reply

Pranav Dar says:
April 05, 2018 at 5:03 pm

Hi Darryl, Thanks for pointing it out. We have replaced it with the open data
with the link.

Reply

Morten Bjoernsvik says:
April 07, 2018 at 10:55 pm

I prefer rest apis so I do not have to download: https://frost.met.no/ - 15K met stations around the globe the oldest data I've found is from 1937 https://data.cityofnewyork.us/browse - I wich my city had the same more than 1000 datasets updated daily.

Reply

Pranav Dar says:
April 09, 2018 at 2:21 pm

Hi Morten, Thanks for sharing these links! These datasets will be useful for our community as well.

Reply

Prashnna says:
April 18, 2018 at 8:33 pm

Those of whom are interested in the Devnagari dataset for character recognition can utilize this (https://github.com/Prasanna1991/DHCD_Dataset) resource. It's bigger than MNIST. P.S. Devnagari script is famous in Indian Subcontinent. (Nepal, India, Bangladesh, etc.)

Reply

Sachin says:
May 06, 2018 at 12:28 am

Thanks for the useful information. Can you please provide the dataset for link prediction.

Reply

Pranav Dar says:
May 07, 2018 at 6:17 pm

Hi Sachin, Can you elaborate on the context around this?

Reply

Vishaal says:
May 08, 2018 at 2:43 am

Thanks, this is a great resource!

Reply

Leo says:
July 28, 2018 at 5:28 am

Bookmarked. Thanks a lot! One note: Wikipedia corpus is definitely bigger than 30MB :)

Reply

Kunal jain says:
November 22, 2018 at 10:59 am

Hi, thank you for the given information....

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

25 Open Datasets for Deep Learning Every Data Scientist Must Work With

Name*

Email*

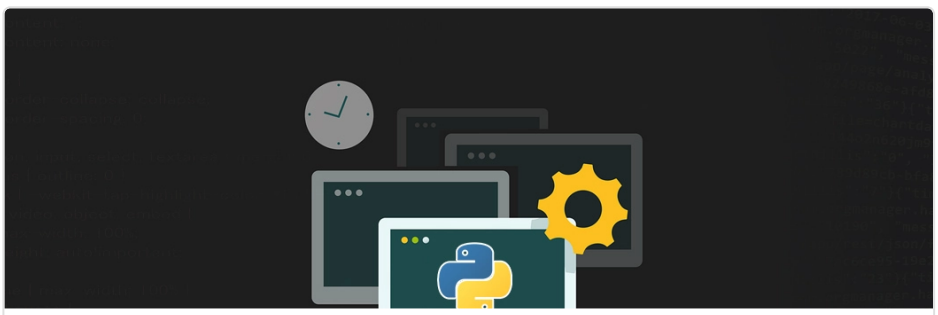
Website

Notify me of follow-up comments by email.

Notify me of new posts by email.

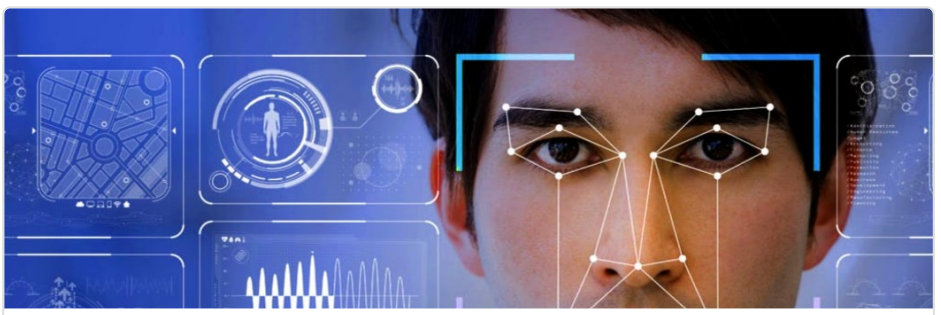
Submit

Top Resources



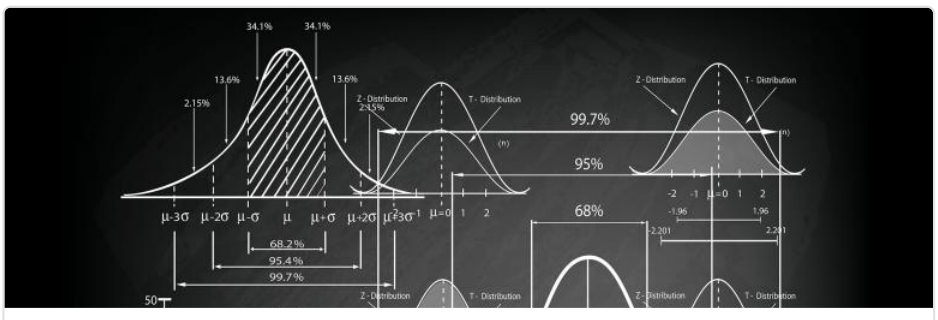
Python Tutorial: Working with CSV file for Data Science

Harika Bonthu - AUG 21, 2021



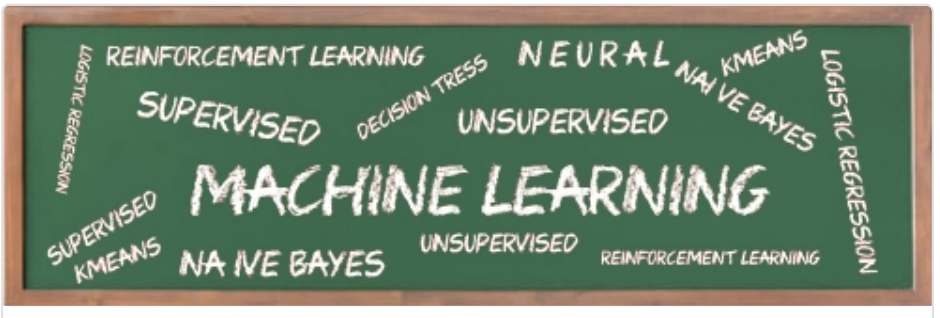
An Introduction to Deepfakes with Only One Source Video

Suvojit Hore - OCT 21, 2021



End to End Statistics for Data Science

Gunjan Agarwal - OCT 29, 2021



Commonly used Machine Learning Algorithms (with Python and R Codes)

Sunil Ray - SEP 09, 2017

25 Open Datasets for Deep Learning Every Data Scientist Must Work With