Crimson Publishers
Wings to the Research

Mini Review

Datasets Finders and Best Public Datasets for Machine Learning and Data Science Applications

Raja Marappan[1]* and S Bhaskaran[2]

[1]Senior Assistant Professor, School of Computing, SASTRA Deemed University, India

[2]Assistant Professor, School of Computing, SASTRA Deemed University, India

ISSN: 2832-4463

***Corresponding author:** Raja Marappan, Senior Assistant Professor, School of Computing, SASTRA Deemed University, India

Submission: April 26, 2022

Published: June 10, 2022

Volume 2 - Issue 1

How to cite this article: Raja Marappan, S Bhaskaran. Datasets Finders and Best Public Datasets for Machine Learning and Data Science Applications. COJ Rob Artificial Intel. 2(1). COJRA. 000530. 2022. DOI: 10.31031/COJRA.2022.02.000530

Copyright@ Raja Marappan, This article is distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract

Recently, there are different public datasets are available for machine learning and data science applications. These datasets are also applied in computer vision, sentiment analysis, clinical data, natural language processing, and others. When anyone shapes the logic of artificial intelligence and data science project, need to write variants of algorithms, and the performance of the model can be evaluated based on the data required for training, and also discuss results of prediction with other methods. Even if anyone is not required to gather particular data collection, the learner can spend more amount of time searching for a better dataset that is required for the implementation of the live project. There are a wide variety of datasets on different categories available in the public domain. The learner must search for the right dataset by spending time. This article is targeted at helping the learners to search for the best available recent datasets in the public domains for artificial intelligence, machine learning, and data science projects.

Keywords: Machine learning; Data science; Datasets; Recommendation; Ratings; Artificial intelligence

Introduction

Machine Learning (ML) is considered the magical model, where one can shuffle the complete information and cast the obtained information into the right predictions. Hence one needs to collect, clean, and merge large amounts of information [1,2]. This article simplifies and gives an exploration of the best online sites at which one can search and get the cumulative datasets for all real-world applications. Datasets are serving like the railways upon which ML methods ride [3-7]. Without these datasets, any ML method will fail to continue in the areas of text mining, categorizing products, and text classification. ML is considered a branch of AI that explores computer science procedures that automatically improve with data processing. The methods of ML use the inputs from historical information to produce efficient prediction measures. ML also plays a central position in world-leading concerns such as Google, Facebook, Amazon, etc. Hence it is necessary to learn the different public datasets before one starts the project.

Dataset Finders

This section explores the best public datasets finders available for data science and ML applications [8-21].

a. Search through Google dataset: Domain: https://datasetsearch.research.google.com/

b. Kaggle dataset: Domain: https://www.kaggle.com/datasets

c. Open-source datasets for UCI ML repository

d. High-quality datasets from CMU Libraries

e. NLP tasks-The Big Bad NLP Database

f. Wikipedia ML Datasets

g. AWS Open Data Registry

h. UCI Machine Learning Repository

Datasets for Real-world Applications

This section explores the best public datasets available for data science and ML applications. The datasets are classified as follows [12-20]:

1. General Datasets: Housing Datasets & Geographic Datasets
2. Machine Learning Datasets: Mall Customers Dataset, IRIS Dataset, MNIST Dataset, Boston Housing Dataset, Fake News Detection Dataset, Wine quality dataset, SOCR data - Heights and Weights Dataset, Titanic Dataset, Credit Card Fraud Detection Dataset
3. Computer Vision Datasets: xView, ImageNet, Kinetics-700, Google's Open Images, Cityscapes Dataset, IMDB-Wiki dataset, Color Detection Dataset, Stanford Dogs Dataset
4. Sentiment Analysis Datasets: Lexicoder Sentiment Dictionary, IMDB reviews, Stanford Sentiment Treebank, Twitter US Airline Sentiment
5. NLP Datasets: The Big Bad NLP Database, HotspotQA Dataset, Amazon Reviews, Rotten Tomatoes Reviews, SMS Spam Collection in English, Enron Email Dataset, Recommender Systems Dataset, UCI Spambase Dataset, IMDB reviews
6. Self-driving (Autonomous Driving) Datasets: Waymo Open Dataset, Berkeley DeepDrive BDD100k, Bosch Small Traffic Light Dataset, LaRa Traffic Light Recognition, WPI datasets, Comma.ai, MIT AGE Lab, LISA: Laboratory for Intelligent & Safe Automobiles, UC San Diego Datasets, Cityscape Dataset
7. Clinical Datasets: MaskedFace-Net, COVID-19 Dataset, MIMIC-III
8. Datasets for Recommender Systems: MovieLens, Jester, Million Song Dataset
9. DataPortals: meta-database with 524 data portals
10. OpenDataSoft: a map with more than 2600 data portals
11. Knoema: home to nearly 3.2-billion time series data of 1040 topics from more than 1200 sources
12. Data.gov: 261,073 sets of the US open government data
13. Eurostat: open data from the EU statistical office
14. Scientific research datasets: Re3data: 2000 research data repositories with flexible search, Harvard Dataverse: 92,839 datasets by the scientific community for the scientific community, Academic torrents: 53.52TB research data aggregated at one place, The Sloan Digital Sky Survey: 3D maps of the Universe
15. Verified datasets from data science communities: DataHub: high-quality datasets shared by data scientists for data scientists, UCI Machine Learning Repository: one of the oldest sources with 488 datasets, data.world: open data community, GitHub: a list of awesome datasets made by the software development community, Kaggle datasets: 25,144 themed datasets on "Facebook for data people", KDnuggets: a comprehensive list of data repositories on a famous data science website, Reddit: datasets and requests of data on a dedicated discussion board
16. Political and social datasets from media outlets: BuzzFeed: datasets and related content by a media company, FiveThirtyEight: datasets from data-driven pieces
17. Finance and economic datasets: Quandl: Alternative Financial and Economic Data, The International Monetary Fund and The World Bank: International Economy Stats
18. Healthcare datasets: World Health Organization: Global Health Records from 194 Countries, The Center for Disease Control (CDC): Searching for data is easy with an online database, Medicare: data from the US health insurance program, The Healthcare Cost and Utilization Project (HCUP): another source with data on healthcare services
19. Travel and transportation datasets: Bureau of Transportation Statistics: the US transportation system in over 260 data tables, Federal Highway Administration: US road transportation data
20. Other sources: Amazon Web Services: free public datasets and paid machine learning tools, Google Public datasets: data analysis with the BigQuery tool in the cloud
21. Earth Dataset: Domain: https://earthdata.nasa.gov/
22. Amazon and Microsoft Datasets, Azure and AWS: Domain AWS: https://registry.opendata.aws/ Domain Azure: https://azure.microsoft.com/en-us/services/open-datasets/catalog/?q=
23. FBI Crime Data Explorer Domain: https://crime-data-explorer.fr.cloud.gov/downloads-and-docs
24. Data World: Domain: https://data.world/
25. CERN Open Data Portal: Domain: http://opendata.cern.ch/
26. Lionbridge AI Datasets: Domain: https://lionbridge.ai/datasets/
27. UCI Machine Learning Repository: Domain: https://archive.ics.uci.edu/ml/index.php
28. Government Datasets for ML: Data USA, EU Open Data Portal, Data.gov, US Healthcare Data, The UK Data Service, School System Finances, The US National Center for Education Statistics
29. Finance & Economics Datasets for ML: American Economic Association (AEA), Quandl, IMF Data, World Bank Open Data, Financial Times Market Data, Google Trends

Copyright © Raja Marappan

30. Image Datasets for Computer Vision: VisualQA, Labelme, ImageNet, Indoor Scene Recognition, Visual Genome, Stanford Dogs Dataset, Google's Open Images, Labelled Faces in the Wild Home, COIL-100, CIFAR-10, Cityscapes, IMDB-Wiki, Fashion MNIST, MS COCO, MPII Human Pose Dataset

31. Sentiment Analysis Datasets for ML: Multi-Domain Sentiment Analysis Dataset, Amazon Product Data, Twitter US Airline Sentiment, IMDB Sentiment, Sentiment140, Stanford Sentiment Treebank, Paper Reviews, Lexicoder Sentiment Dictionary, Sentiment Lexicons for 81 Languages, Opin-Rank Review Dataset

32. NLP Datasets: Enron Dataset, UCI's Spambase, Amazon Reviews, Yelp Reviews, Google Books Ngrams, SMS Spam Collection in English, Jeopardy, Gutenberg eBooks List, Blogger Corpus, Wikipedia Links Data

33. Datasets for Autonomous Vehicles: Berkeley DeepDrive BDD100K, Comma.ai, Oxford's Robotic Car, LISA, Cityscapes Dataset, Baidu Apolloscapes, Landmarks, Landmarks-v2, PandaSet, nuScenes, Open Images V5, Waymo Open Dataset

34. Recommendation and Ratings Public Data Sets For ML: Movies Recommendation, Music Recommendation, Books Recommendation, Food Recommendation, Merchandise Recommendation, Healthcare Recommendation, Dating Recommendation, Scholarly Paper Recommendation

Conclusion & Future Work

This article gives an overview of different public datasets available for ML and data science applications. The best public datasets finders available for data science and ML applications are also listed. The required datasets available in the public domains for data science and ML applications are classified with several examples. In the future, these public datasets can be applied with soft computing and approximation algorithms for solving the different real-world applications [22-41].

References

1. 50 Best free datasets for machine learning, Lionbridge AI, Massachusetts, USA.
2. Google cloud public datasets, Google.
3. Machine learning and AI datasets, Carnegie Mellon University, USA.
4. Big data and AI: 30 amazing and free public data sources, Forbes, New Jersey, USA.
5. Awesome autonomous vehicles datasets, Github, California, USA.
6. Fueling the gold rush, The greatest public datasets for AI, StartupGrind.
7. Places to find free datasets for data science projects, Dataquest.
8. The best datasets for natural language processing, Gengo AI.
9. Awesome public datasets, Github, California, USA.
10. StatLib datasets archive, Carnegie Mellon, USA.
11. Institutional research and analysis, Common Datasets.
12. Datasets and project suggestions, Andrew W Moore.
13. Datasets, Machine Learning Repository, MIT, USA
14. Datasets, MIT Lincoln Laboratory, USA.
15. Stanford large network dataset collection, Stanford University, USA.
16. Stanford common dataset, Stanford University, USA.
17. Datalab, UC Berkeley.
18. Exploring datasets, Data Science at Berkeley.
19. DeepDrive, UC Berkeley.
20. Machine learning datasets and project ideas-work on real-time data science projects, Data Flair.
21. Cabani A, Hammoudi K, Benhabiles H, Melkemi M (2020) MaskedFace-Net-A dataset of correctly/incorrectly masked face images in the context of COVID-19, Smart Health 19: 100144.
22. Marappan R, Sethumadhavan G (2013) A new genetic algorithm for graph coloring. 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, Seoul, Korea (South), pp. 49-54.
23. Sethumadhavan G, Marappan R (2013) A genetic algorithm for graph coloring using single parent conflict gene crossover and mutation with conflict gene removal procedure. 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, India, pp. 1-6.
24. Marappan R, Sethumadhavan G (2015) Solving graph coloring problem for large graphs. Global Journal of Pure and Applied Mathematics 11(4): 2487-2494.
25. Marappan R, Sethumadhavan G (2016) Solving channel allocation problem using new genetic algorithm with clique partitioning method. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, India, pp. 1-4.
26. Marappan R, Sethumadhavan G (2015) Solution to graph coloring problem using evolutionary optimization through symmetry-breaking approach. International Journal of Applied Engineering Research 10(10): 26573-26580.
27. Marappan R, Sethumadhavan G (2016) Solution to graph coloring problem using divide and conquer based genetic method. 2016 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, pp. 1-5.
28. Marappan R, Sethumadhavan G (2015) Solution to graph coloring problem using heuristics and recursive backtracking. International Journal of Applied Engineering Research 10(10): 25939-25944.
29. Marappan R, Sethumadhavan G, Srihari RK (2016) New approximation algorithms for solving graph coloring problem-An experimental approach. Perspectives in Science 8: 384-387.
30. Marappan R, Sethumadhavan G, Harimoorthy U (2016) Solving channel allocation problem using new genetic operators- An experimental approach. Perspectives in Science, Volume 8: 409-411.
31. Marappan R, Sethumadhavan G (2016) Divide and conquer based genetic method for solving channel allocation. 2016 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, pp. 1-5.
32. Marappan R, Sethumadhavan G (2016) Solving fixed channel allocation using hybrid evolutionary method. MATEC Web of Conferences 57: 02015.
33. Marappan R, Sethumadhavan G (2018) Solution to graph coloring using genetic and tabu search procedures. Arab J Sci Eng 43: 525-542.
34. Marappan R, Sethumadhavan G (2020) Complexity analysis and stochastic convergence of some well-known evolutionary operators for solving graph coloring problem. Mathematics 8: 303.

Copyright © Raja Marappan

35. Bhaskaran S, Marappan R, Santhi B (2020) Design and comparative analysis of new personalized recommender algorithms with specific features for large scale datasets. Mathematics 8: 1106.

36. Marappan R, Sethumadhavan G (2021) Solving graph coloring problem using divide and conquer-based turbulent particle swarm optimization. Arab J Sci Eng.

37. Bhaskaran S, Marappan R, Santhi B (2021) Design and analysis of a cluster-based intelligent hybrid recommendation system for e-learning applications. Mathematics 9(2): 197.

38. Bhaskaran S, Marappan R (2021) Design and analysis of an efficient machine learning based hybrid recommendation system with enhanced density-based spatial clustering for digital e-learning applications. Complex Intell Syst.

39. Balakrishnan S, Suresh T, Marappan R (2021) Analysis of recent trends in solving np problems with new research directions using evolutionary methods. International Journal of Research Publication and Reviews 2(8): 1429-1435.

40. Balakrishnan S, Suresh T, Marappan R (2021) A new multi-objective evolutionary approach to graph coloring and channel allocation problems. Journal of Applied Mathematics and Computation, 5(4): 252-263.

41. Raja Marappan (2021) A new multi-objective optimization in solving graph coloring and wireless networks channels allocation problems. Int J Advanced Networking and Applications 13(2): 4891-4895.

For possible submissions Click below:

Submit Article

Copyright © Raja Marappan