Analytics Vidhya

TOP 10 GitHub Repositories for Data Science

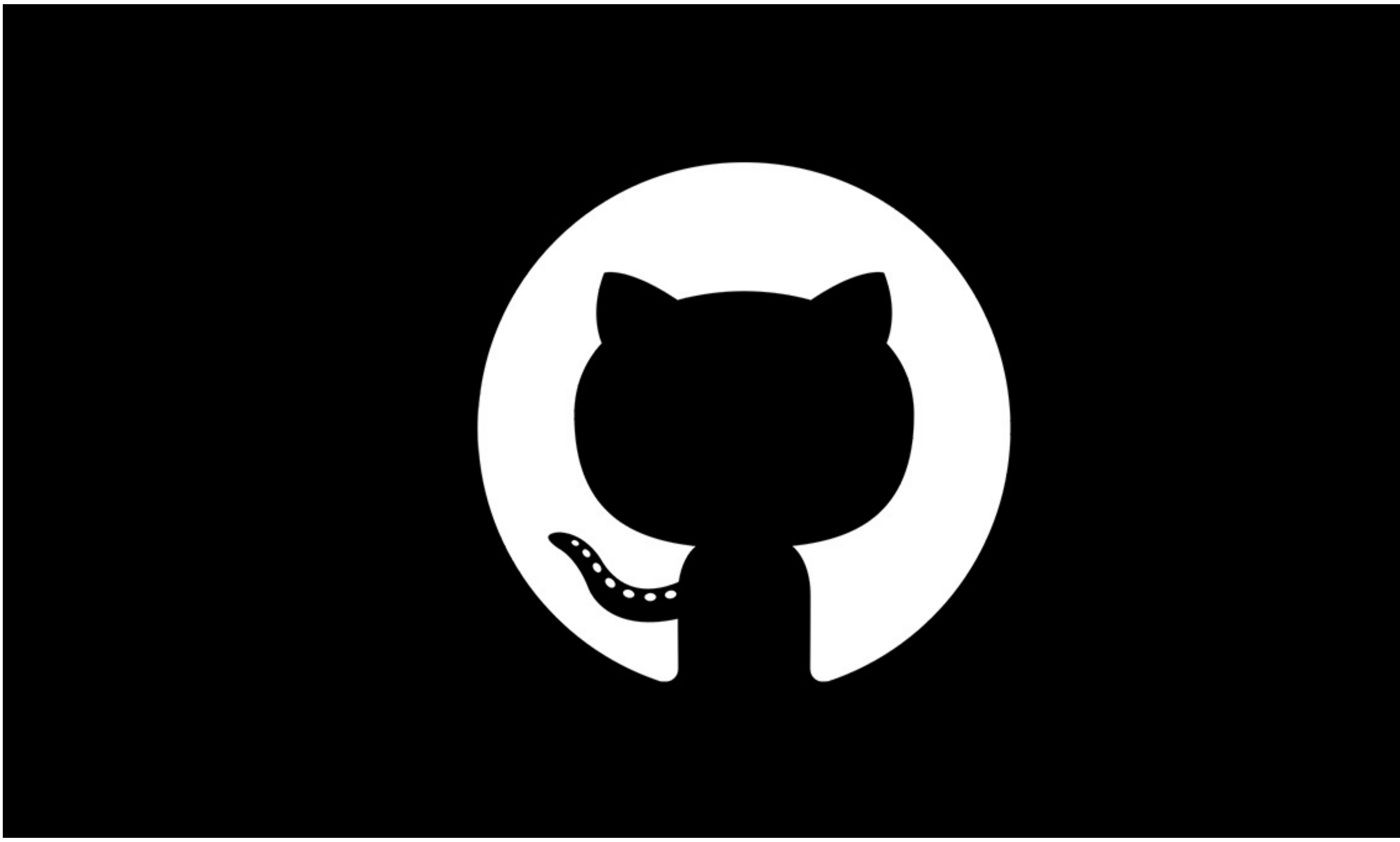
DATA SCIENCE DATASETS GITHUB PROJECT

Introduction

Data science is a collaborative scientific field of computing that has grown many folds in recent years and has become the powerhouse behind the business decisions made by organizations in today's time, be it the FAANG's or early-stage startups.

As the field has grown, so have the number of individuals pursuing this domain and the learning resources available on the internet. The premier resource to learn data science is GitHub among all these resources.

What is GitHub?



As the word itself, GitHub suggests a hub for over 73 million coders and developers to host and share codes in a cooperative and collaborative environment. It provides several features like access control, version control and continuous integration for every project and is the most prominent source code host globally with over 28 million public repositories. I have compiled the top 10 repositories for learning data science out of these.

To know more about GitHub, read here.

FREECODECAMP (337K ★)

Freecodecamp is free to learn online coding community with a speciality in various domains. It provides several certifications on different code profiles, including Data Visualization Certification, Data Analysis with Python Certification and Machine Learning with Python Certification. Freecodecamp community also has a forum where users can get programming help and feedback on their projects. They also have a Youtube channel that contains free courses on Python, SQL, Machine Learning and many more.

TENSORFLOW (161K ★)

TensorFlow

TensorFlow is an open-source framework for Machine learning and Artificial Intelligence developed by Google Brain Team. The GitHub repository contains various resources to learn and enhance the TensorFlow and Machine Learn skills.

You can learn more about TensorFlow through TensorFlow tutorials. These tutorials are written in Jupyter notebooks and can be run directly on Google Colab requiring no setup.

It also provides state of the art models for Machine Learning in domains such as computer vision, NLP, and recommendation systems. They are highly optimized and efficient at the task they are designed to do, which enables them to use them directly and generate highly accurate results on their datasets.

THE ALGORITHMS (126K ★)

This GitHub repository contains various algorithms coded exclusively in Python. It enlists a collection of codes on domains such as Machine learning, Neural Networks, Digital Image Processing and Computer Vision.

The Machine Learning sub-repository provides codes on several regression techniques such as linear and polynomial regression. They are usually used in predictive analysis for continuous data and are very useful for problems pertaining to stock price or house prediction. It also contains classification methods such as logistic regression and multi-layer perceptron used to predict data containing discrete values(where data is divided into many classes).

The neural network repository contains codes on backpropagation which deals with updating weights in neural network architecture, Convolutional Neural Network provides the machine human like ability to distinguish between different classes of images. One of the most common applications of the CNN architecture is “”oogle Lens””

The digital image repository contains codes on edge detection such as Canny edge detection. These types of techniques are more often used to isolate the edges in an environment capture. One of the most known applications is autonomous cars which rely on the same for determining the road linings.

The computer vision repository contains codes for pooling, a feature of CCNN’sthat is used to extract the highest rated features in an image for classification.

AWESOME MACHINE LEARNING (52.2K ★)

The above given GitHub repository provides an organized list of machine learning libraries, frameworks and tools in almost all the languages available. As most of Machine Learning development is done on Python, practitioners without Python as their background may find it difficult to adapt to it. So this makes this repository even more valuable as it transcends all the languages and promotes a collective development environment for Machine Learning.

In python, the libraries are provided on the following domains:

- Computer Vision
- Natural Language Processing
- General-Purpose Machine Learning
- Data Analysis / Data Visualization
- Neural Networks

Further elaborating, Computer vision libraries include scikit-image, scikit-opt, face_recognition, neural dream and many more, NLP libraries include CLTK and NLTK which helps us to build models that are able to understand human language data, Machine learning libraries include scikit learn, pattern and prophet which was developed by Facebook and is one of the best models for time series data prediction, Data Visualization and analysis libraries include pandas, numpy and many more which are really helpful in modelling and transforming our datasets and finally neural network libraries include neural_talk, nn_builder which can build neural networks in one line!!.

DATA SCIENCE I-PYTHON NOTEBOOKS (22.1K ★)

The above-given repository contains python notebooks on almost every aspect of machine learning, data engineering and data augmentation. It has data science python notebooks for Deep learning libraries and frameworks (TensorFlow, Theano, Caffe, Keras), scikit-learn, big data notebooks on Spark, Hadoop MapReduce, HDFS, data visualization notebooks on matplotlib, and data transformation notebooks on pandas, NumPy, SciPy.

Among them, one of the most popular libraries is scikit-learn which contains notebooks for several machine learning algorithms such as K-Nearest Neighbors, Support Vector Machines, Random Forest, K-Means and principal components analysis.

Through the pandas i-notebooks one can learn techniques such as data indexing, merging joining, aggregation and filling in missing values. All of this comes under data cleaning and preparation and is the most important part of the data analysis pipeline. In fact, without data cleaning and augmentation no amount of analysis thorough different algorithms would yield any valuable or sensible results.

Through Matplotlib notebooks people can learn about creating user-friendly bar graphs and charts which are really helpful in depicting analysis results in a user-friendly way.

HOMEMADE MACHINE LEARNING (18.6K ★)

This repository contains instances of the most used and widely used machine learning codes and algorithms implemented using Python explained along with the mathematics and logic working behind them. Also, each algorithm is explained through Jupyter notebook'sinteractive environment. The codes are not only run on a training set for data analysis but also the mathematics is explained which makes it one of the best resources to strengthen one's basics.

For supervised learning it provides assistance for regression and classification techniques by explaining the mathematics behind linear regression, logistic regression providing the code for it and running it on Jupyter notebook.

For unsupervised learning, it provides codes for clustering which is used in problems such as customer segmentation. In clustering, we split the training examples into different clusters based on columns of data whose legends are not known to us.

For neural network it provides an explanation on multi-layer perceptron, working of activation functions, cost functions, loss functions and gradient descent.

AWESOME DATA SCIENCE (17.6K ★)

This GitHub repository is very important for those who want to understand the basics of data science and Machine Learning. It takes you from answering your elementary questions such as “”hat is data science”” “”hy we need to use it”” “”hat are its applications””and brings you to a position where you would be well versed with the basics of data science.

It also contains a curated list of MMOOC’s which is in my opinion one of the best ways one can gain knowledge in this domain.

It also contains several tutorials and free courses for you to start your data science journey.

It also contains a list of libraries used for deep learning, machine learning, tensorflow, Keras which are extensively used in each and every code you would come across in data science.

Also, you can find top journals, publications and magazines on data science and Big Data which is really helpful to remain up to date with the latest developments in the field.

For those who prefer listening over reading, you are in luck as it contains an exclusive list of podcasts and YouTube Channels on several data science topics such as AI, Big data and data engineering.

You can also follow up on reading the most popular books on data science and exchange your ideas and follow the most prominent bloggers.

DEEP LEARNING DRIZZLE (9.7K ★)

As the name suggests deep learning drizzle is a GitHub repository dedicated to deep learning algorithms. It provides resources such as lecture slides of the most prominent universities of the world and their YouTube lectures on several domains such as:

Deep Neural Networks

Machine Learning Fundamentals

Natural Language Processing

Optimization for Machine Learning

General Machine Learning

Modern Computer Vision and many more.

These resources are highly valued and followed by millions of people across the globe. Thus they are bound to provide you with extensive knowledge of deep neural architecture and machine learning in general.

500 AI-ML PROJECTS (7.4K ★)

One of the major parts of learning any field be it data science, AI or any other is to have hands-on knowledge, to have practical experience. Most of the people studying or pursuing their interests in this field come across an opportunity to create projects on data science. So, this repository provides you with one of the most important lists containing over 500 projects on machine learning, NLP, AI along code. This is really helpful for those who want hands-on knowledge or want to create projects for their resume.

INTERACTIVE TOOLS (1.4K ★)

This repository contains interactive tools for deep learning, machine learning along with an explanation of the math behind it. It is really intuitive and a new way to understand and comprehend the complex nature of these algorithms. Their work is depicted through videos which help to see how they are converting and analyzing the data in real-time.

Take for instance the CNN Explainer, which is an interactive video description that explains the working of a convolutional network. And for each of these examples the code, demo video and official paper are given.

ENDNOTES

Through the medium of this article, we have journeyed through the catalogue of the best GitHub repositories on theInternett. From free resources to interactive tools and from free courses to awesome codes, we have gone through some amazing work developed and provided to us for the taking. I am sure that, even if one imbibes a part of the assortment of these resources, she/he can excel and reach newer heights at their career in data science.

REFERENCES

Image 1- https://rb.gy/9zxi9v

Image 2- https://rb.gy/cwjc93

Image 3- https://rb.gy/x1h5s9

Image 4- https://rb.gy/ecgxcf

Image 5- https://rb.gy/qccyul

Image 6- https://rb.gy/29ihlg

Image 7- https://rb.gy/rnbhvt

Image 8- https://rb.gy/x9vmlq

Image 9- https://rb.gy/shm6ln

Article Url - https://www.analyticsvidhya.com/blog/2022/01/top-10-github-repositories-for-data-science/

Ayushi Gupta