Understanding K-means Clustering in Machine Learning

Education Ecosystem (LEDU) · Follow
Published in Towards Data Science
5 min read · Sep 12, 2018

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

AndreyBu, who has more than 5 years of machine learning experience and currently teaches people his skills, says that "the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset."

A cluster refers to a collection of data points aggregated together because of certain similarities.

You'll define a target number *k*, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The *'means'* in the K-means refers to averaging of the data; that is, finding the centroid.

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

How the K-means algorithm works

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

K-means algorithm example problem

Let’s see the steps on how the K-means machine learning algorithm works using the Python programming language.

We’ll use the Scikit-learn library and some random data to illustrate a K-means clustering simple explanation.

Step 1: Import libraries

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

%matplotlib inline
```

As you can see from the above code, we’ll import the following libraries in our project:

- Pandas for reading and writing spreadsheets
- Numpy for carrying out efficient computations
- Matplotlib for visualization of data

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

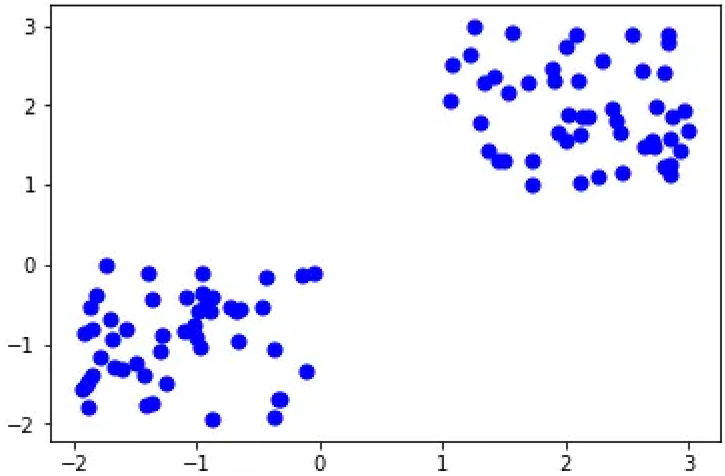
Step 2: Generate random data

Here is the code for generating some random data in a two-dimensional space:

```
X= -2 * np.random.rand(100,2)

X1 = 1 + 2 * np.random.rand(50,2)

X[50:100, :] = X1

plt.scatter(X[ : , 0], X[ :, 1], s = 50, c = ‘b’)

plt.show()
```

A total of 100 data points has been generated and divided into two groups, of 50 points each.

Here is how the data is displayed on a two-dimensional space:



Step 3: Use Scikit-Learn

We’ll use some of the available functions in the Scikit-learn library to process the randomly generated data.

Here is the code:

```
from sklearn.cluster import KMeans

Kmean = KMeans(n_clusters=2)
```

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

```
Kmean.fit(X)
```

In this case, we arbitrarily gave *k* (n_clusters) an arbitrary value of two.

Here is the output of the K-means parameters we get if we run the code:

```
KMeans(algorithm=’auto’, copy_x=True, init=’k-means++’,
max_iter=300
 n_clusters=2, n_init=10, n_jobs=1, precompute_distances=’auto’,
 random_state=None, tol=0.0001, verbose=0)
```

Step 4: Finding the centroid

Here is the code for finding the center of the clusters:

```
Kmean.cluster_centers_
```

Here is the result of the value of the centroids:

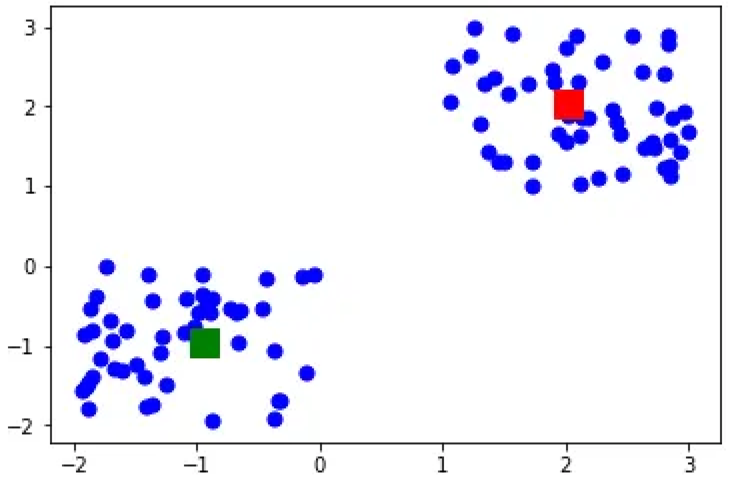
```
array([[-0.94665068, -0.97138368],
 [ 2.01559419, 2.02597093]])
```

Let’s display the cluster centroids (using green and red color).

```
plt.scatter(X[ : , 0], X[ : , 1], s =50, c=’b’)

plt.scatter(-0.94665068, -0.97138368, s=200, c=’g’, marker=’s’)

plt.scatter(2.01559419, 2.02597093, s=200, c=’r’, marker=’s’)

plt.show()
```

Here is the output:

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF



Step 5: Testing the algorithm

Here is the code for getting the labels property of the K-means clustering example dataset; that is, how the data points are categorized into the two clusters.

```
Kmean.labels_
```

Here is the result of running the above K-means algorithm code:

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```

As you can see above, 50 data points belong to the **0** cluster while the rest belong to the **1** cluster.

For example, let’s use the code below for predicting the cluster of a data point:

```
sample_test=np.array([-3.0,-3.0])

second_test=sample_test.reshape(1, -1)

Kmean.predict(second_test)
```

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

Here is the result:

```
array([0])
```

It shows that the test data point belongs to the **0** (green centroid) cluster.

Wrapping up

Here is the entire K-means clustering algorithm code in Python:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
%matplotlib inline
X= -2 * np.random.rand(100,2)
X1 = 1 + 2 * np.random.rand(50,2)
X[50:100, :] = X1
plt.scatter(X[ : , 0], X[ :, 1], s = 50, c = ‘b’)
plt.show()
from sklearn.cluster import KMeans
Kmean = KMeans(n_clusters=2)
Kmean.fit(X)
Kmean.cluster_centers_
plt.scatter(X[ : , 0], X[ : , 1], s =50, c=’b’)
plt.scatter(-0.94665068, -0.97138368, s=200, c=’g’, marker=’s’)
plt.scatter(2.01559419, 2.02597093, s=200, c=’r’, marker=’s’)
plt.show()
Kmean.labels_
sample_test=np.array([-3.0,-3.0])
second_test=sample_test.reshape(1, -1)
Kmean.predict(second_test)
```

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

K-means clustering is an extensively used technique for data cluster analysis.

It is easy to understand, especially if you accelerate your learning using a K-means clustering tutorial. Furthermore, it delivers training results quickly.

However, its performance is usually not as competitive as those of the other sophisticated clustering techniques because slight variations in the data could lead to high variance.

Furthermore, clusters are assumed to be spherical and evenly sized, something which may reduce the accuracy of the K-means clustering Python results.

What’s your experience with K-means clustering in machine learning?

Please share your comments below.

Machine Learning

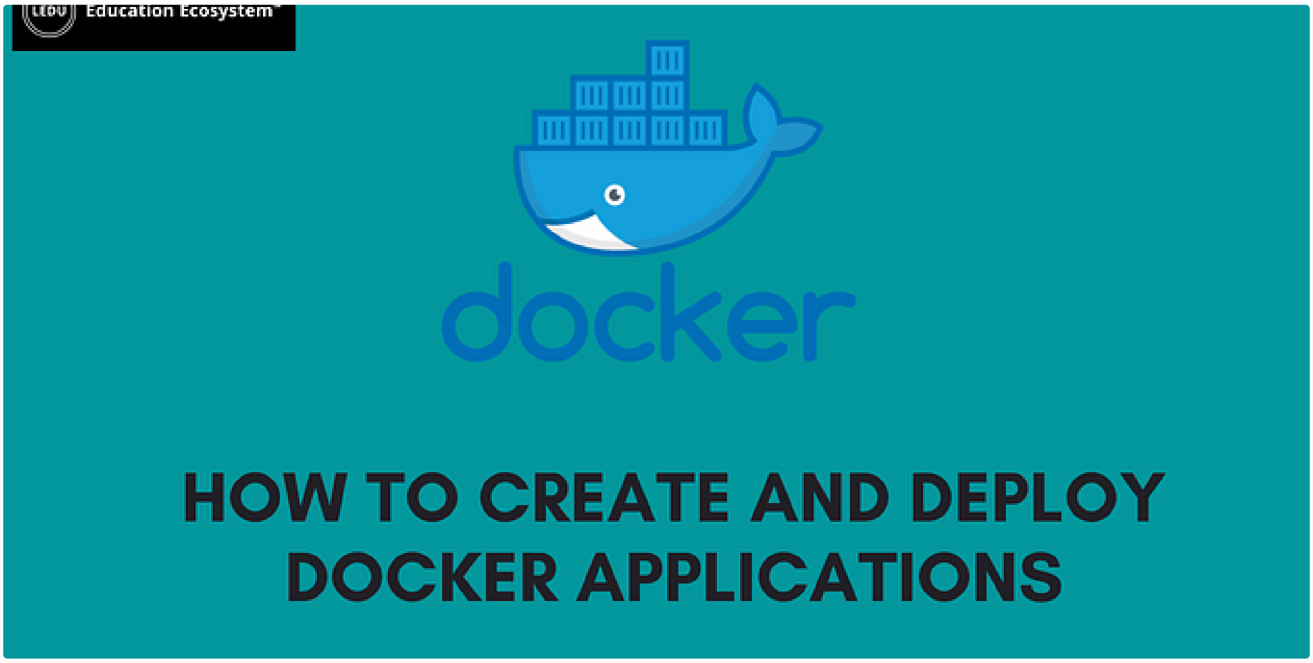
Written by Education Ecosystem (LEDU)

5.6K Followers · Writer for Towards Data Science

Education Ecosystem (LEDU) is a decentralized project-based learning platform that teaches people how to build tech products, https://www.educationecosystem.com

More from Education Ecosystem (LEDU) and Towards Data Science

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

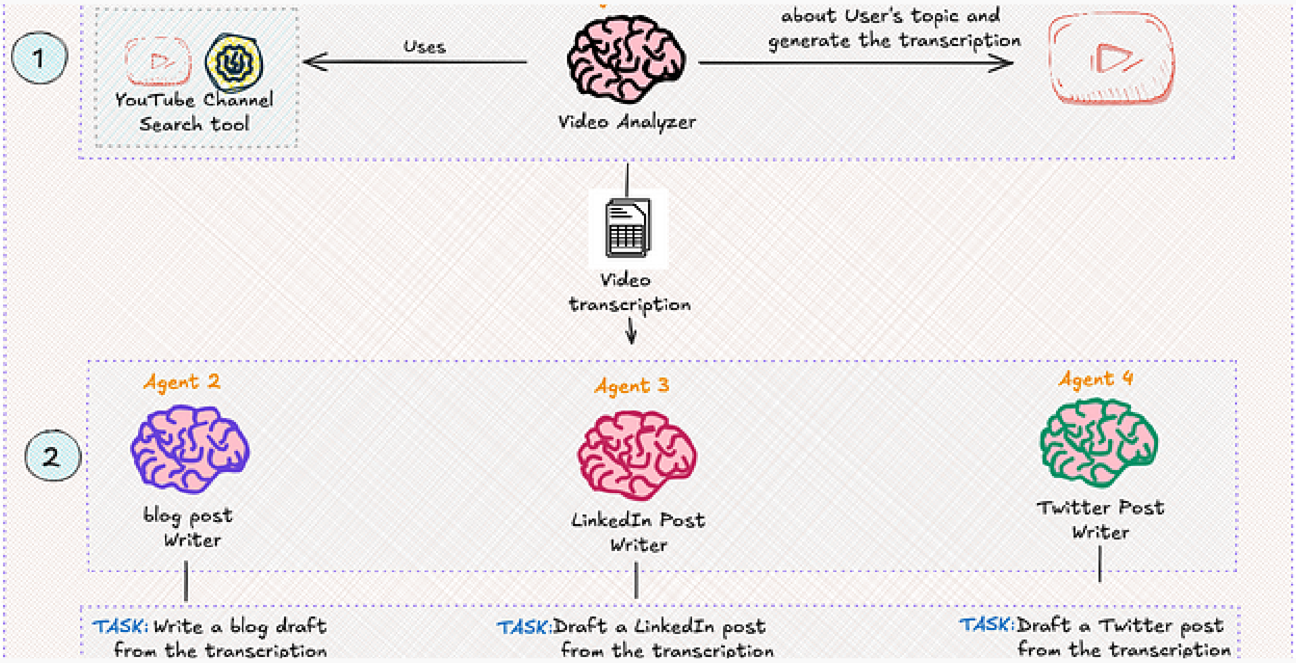


Education Ecosystem (LEDU) in Geek Culture

How To Create And Deploy Docker Applications

A brief overview of Docker

Jul 16, 2022 17


1
YouTube Channel Search tool
Uses
Video Analyzer
about User's topic and generate the transcription
Video transcription
2
Agent 2
blog post Writer
Agent 3
LinkedIn Post Writer
Agent 4
Twitter Post Writer
TASK: Write a blog draft from the transcription
TASK: Draft a LinkedIn post from the transcription
TASK: Draft a Twitter post from the transcription

Zoumana Keita in Towards Data Science

AI Agents — From Concepts to Practical Implementation in Python

This will change the way you think about AI and its capabilities

Aug 12 1.1K 13

Ahmed Besbes in Towards Data Science

What Nobody Tells You About RAGs

A deep dive into why RAG doesn't always work as expected: an overview of the business value, the data, and the technology behind it.

Aug 23 1.5K 22

Education Ecosystem (LEDU)

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

Get Started With Linux: A Beginner's Guide

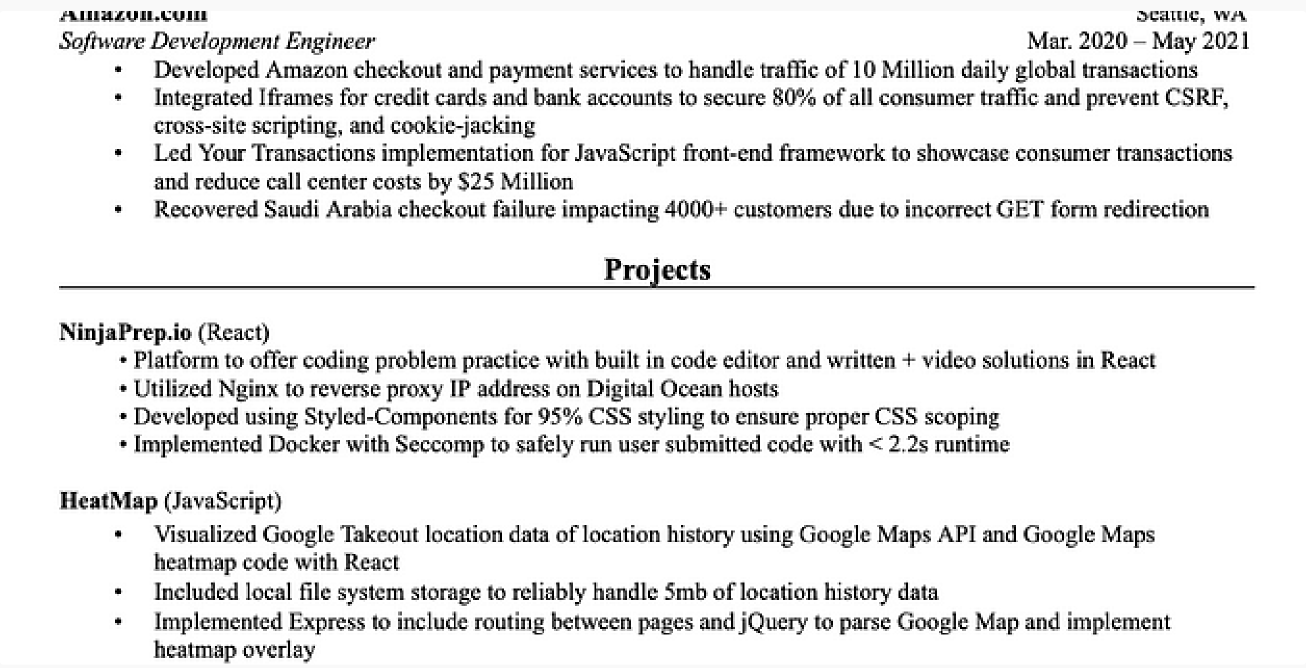
Are you looking to learn Linux? If you do, you have come to the right place.

Aug 10, 2019 · 581 · 1

See all from Education Ecosystem (LEDU)

See all from Towards Data Science

Recommended from Medium



Amazon.com Seattle, WA

Software Development Engineer Mar. 2020 – May 2021

- Developed Amazon checkout and payment services to handle traffic of 10 Million daily global transactions
- Integrated Iframes for credit cards and bank accounts to secure 80% of all consumer traffic and prevent CSRF, cross-site scripting, and cookie-jacking
- Led Your Transactions implementation for JavaScript front-end framework to showcase consumer transactions and reduce call center costs by $25 Million
- Recovered Saudi Arabia checkout failure impacting 4000+ customers due to incorrect GET form redirection

Projects

NinjaPrep.io (React)

- Platform to offer coding problem practice with built in code editor and written + video solutions in React
- Utilized Nginx to reverse proxy IP address on Digital Ocean hosts
- Developed using Styled-Components for 95% CSS styling to ensure proper CSS scoping
- Implemented Docker with Seccomp to safely run user submitted code with < 2.2s runtime

HeatMap (JavaScript)

- Visualized Google Takeout location data of location history using Google Maps API and Google Maps heatmap code with React
- Included local file system storage to reliably handle 5mb of location history data
- Implemented Express to include routing between pages and jQuery to parse Google Map and implement heatmap overlay

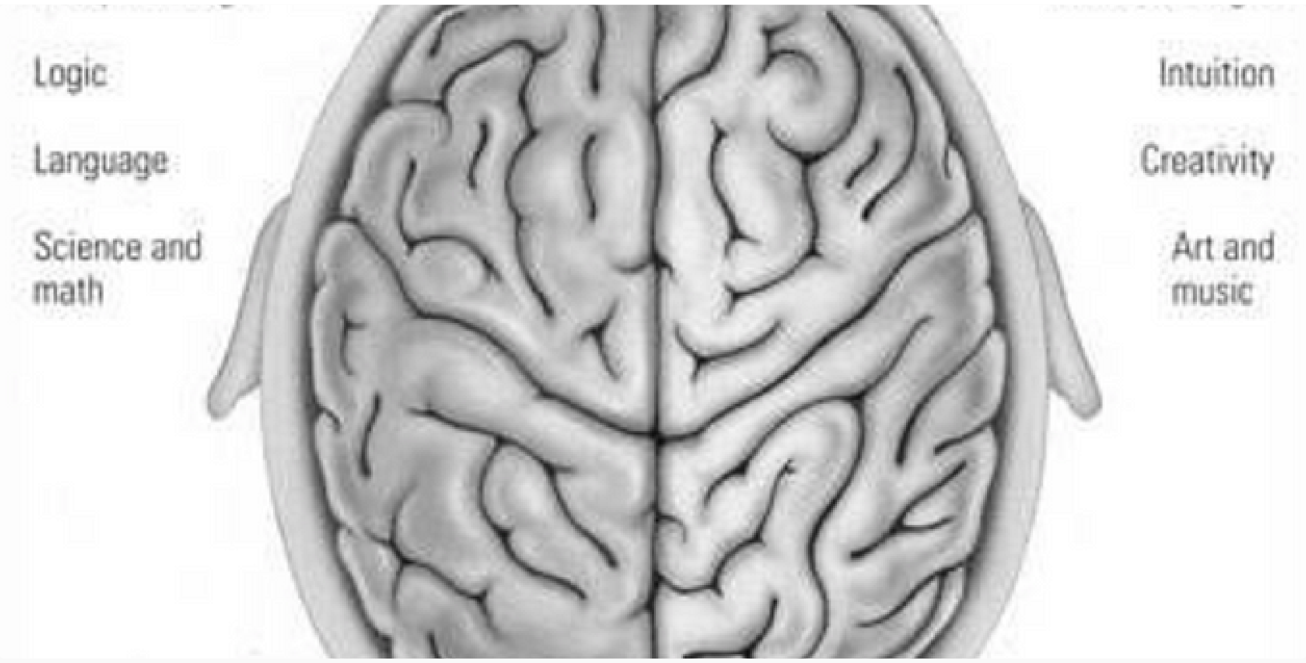
Alexander Nguyen in Level Up Coding

The resume that got a software engineer a $300,000 job at Google.

1-page. Well-formatted.

Jun 1 · 20K · 384

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF



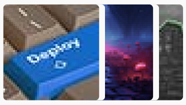
Sufyan Maan, M.Eng

What Happens When You Start Reading Every Day

Think before you speak. Read before you think.—Fran Lebowitz

Mar 12 33K 808

Lists

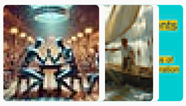


Predictive Modeling w/ Python

20 stories · 1500 saves

Practical Guides to Machine Learning

10 stories · 1831 saves



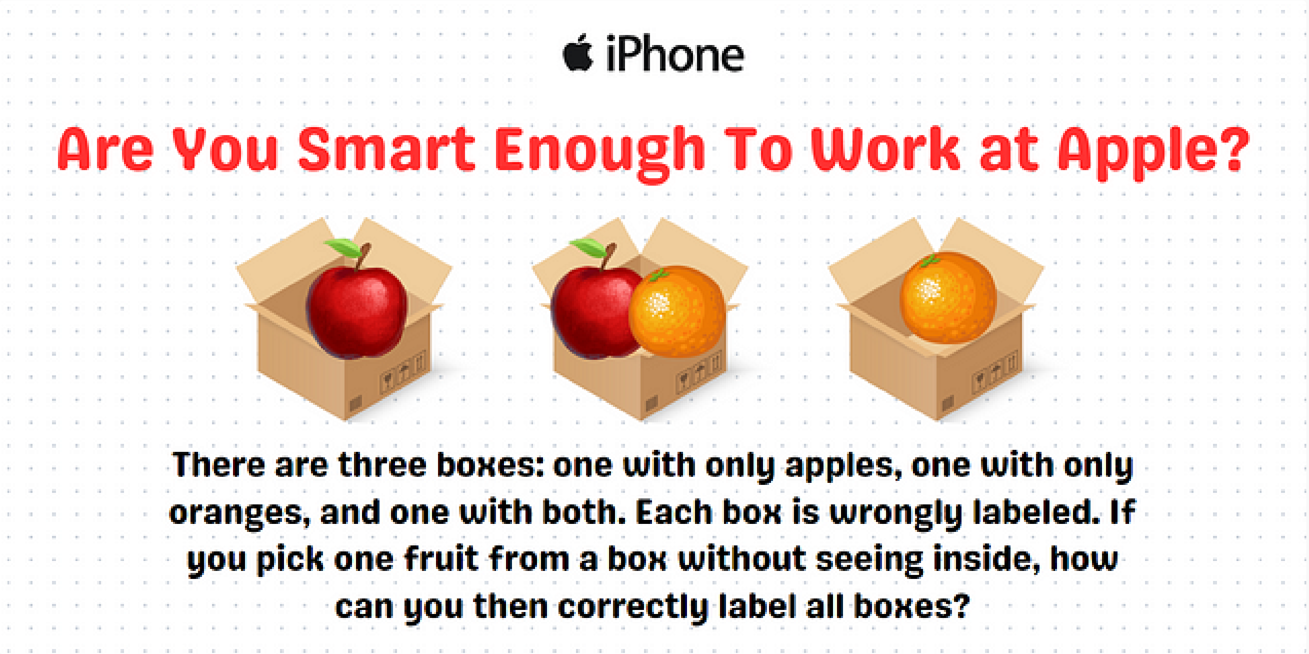
Natural Language Processing

1687 stories · 1254 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 455 saves

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF



Are You Smart Enough To Work at Apple?

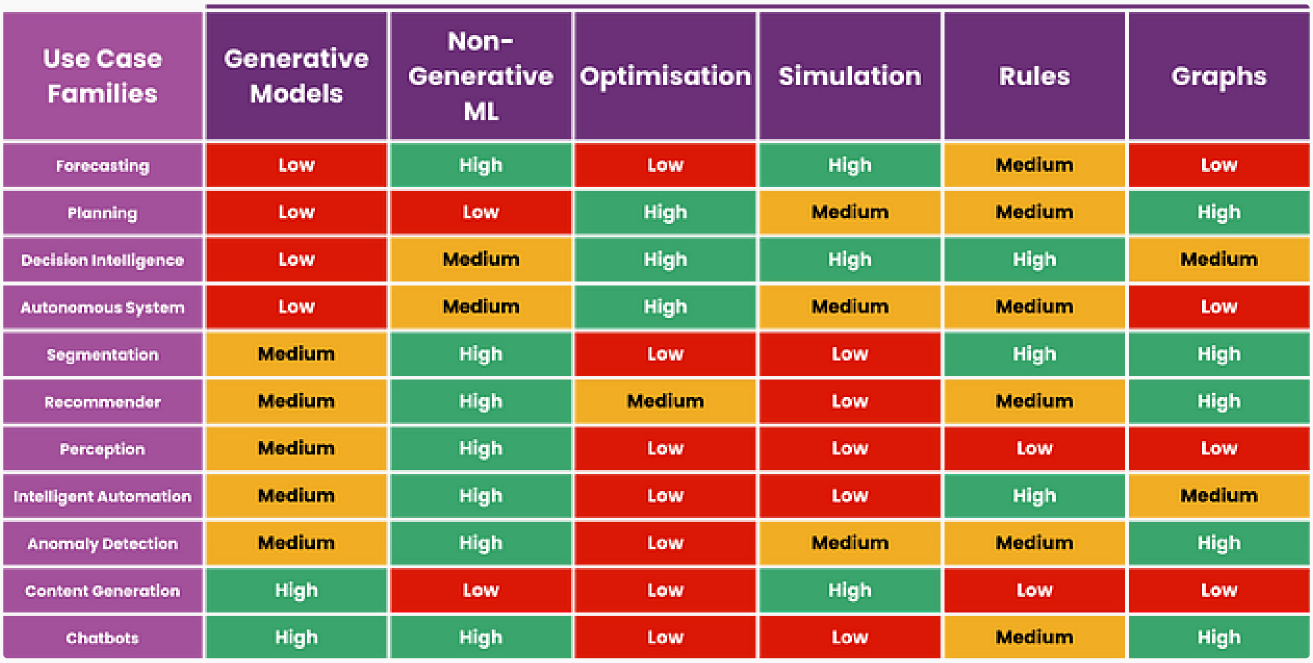
There are three boxes: one with only apples, one with only oranges, and one with both. Each box is wrongly labeled. If you pick one fruit from a box without seeing inside, how can you then correctly label all boxes?

Bella L in ILLUMINATION

Can You Pass This Apple-Orange Interview At Apple 🍎?

The iPhone Company's Interview Question

✦ Mar 13 👏 8.4K 💬 242



Use Case Families	Generative Models	Non-Generative ML	Optimisation	Simulation	Rules	Graphs
Forecasting	Low	High	Low	High	Medium	Low
Planning	Low	Low	High	Medium	Medium	High
Decision Intelligence	Low	Medium	High	High	High	Medium
Autonomous System	Low	Medium	High	Medium	Medium	Low
Segmentation	Medium	High	Low	Low	High	High
Recommender	Medium	High	Medium	Low	Medium	High
Perception	Medium	High	Low	Low	Low	Low
Intelligent Automation	Medium	High	Low	Low	High	Medium
Anomaly Detection	Medium	High	Low	Medium	Medium	High
Content Generation	High	Low	Low	High	Low	Low
Chatbots	High	High	Low	Low	Medium	High

Christopher Tao in Towards AI

Do Not Use LLM or Generative AI For These Use Cases

Choose correct AI techniques for the right use case families

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

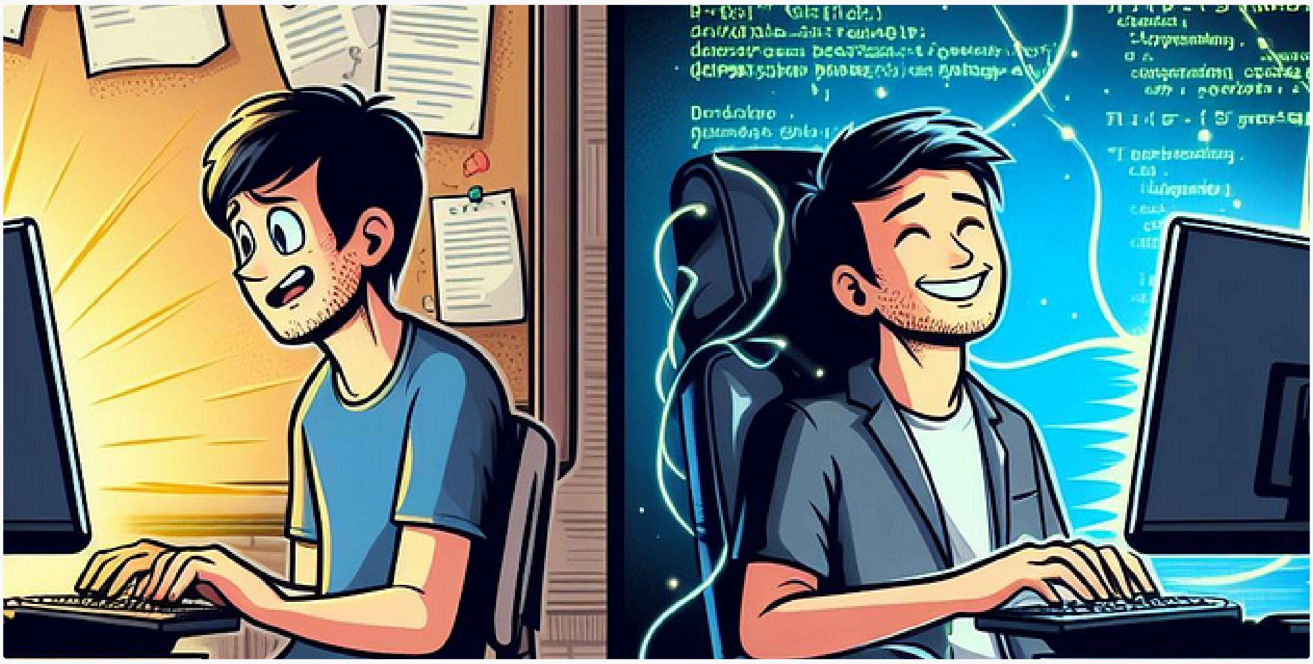
✦ Aug 10 👏 2.9K 💬 35

Hazel Paradise

How I Create Passive Income With No Money

many ways to start a passive income today

Mar 27 👏 22K 💬 671



Abhay Parashar in The Pythoneers

17 Mindblowing Python Automation Scripts I Use Everyday

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF

Scripts That Increased My Productivity and Performance

✦ Aug 25 7.4K 71

See more recommendations

Explore our developer-friendly HTML to PDF API Printed using PDFCrowd HTML to PDF