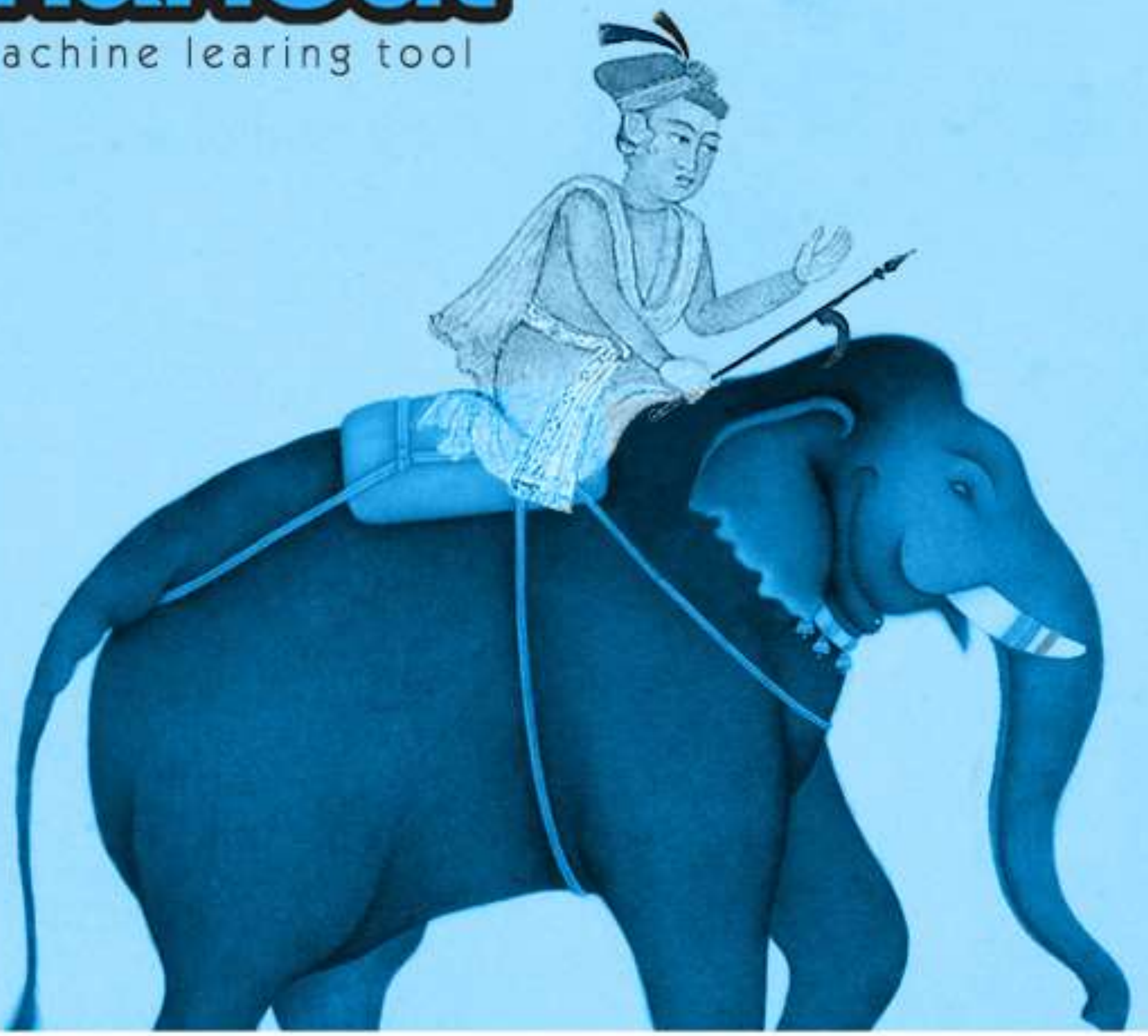
mahout

machine learing tool



tutorialspoint

SIMPLY EASY LEARNING

www.tutorialspoint.com

https://www.facebook.com/tutorialspointindia

https://twitter.com/tutorialspoint

About this Tutorial

Apache Mahout is an open source project that is primarily used in producing scalable machine learning algorithms. This brief tutorial provides a quick introduction to Apache Mahout and explains how it can be applied to make recommendations and organize documents in more useable clusters.

Audience

This tutorial has been prepared for professionals aspiring to learn the basics of Mahout and develop applications involving machine learning techniques such as recommendation, classification, and clustering.

Prerequisites

Before you start proceeding with this tutorial, we assume that you have prior exposure to Core Java, Hadoop, and any of the Linux operating system flavors.

Copyright & Disclaimer

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at contact@tutorialspoint.com

Table of Contents

1. MAHOUT —INTRODUCTION

We are living in a day and age where information is available in abundance. The information overload has scaled to such heights that sometimes it becomes difficult to manage our little mailboxes! Imagine the volume of data and records some of the popular websites (the likes of Facebook, Twitter, and Youtube) have to collect and manage on a daily basis. It is not uncommon even for lesser known websites to receive huge amounts of information in bulk.

Normally we fall back on data mining algorithms to analyze bulk data to identify trends and draw conclusions. However, no data mining algorithm can be efficient enough to process very large datasets and provide outcomes in quick time, unless the computational tasks are run on multiple machines distributed over the cloud.

We now have new frameworks that allow us to break down a computation task into multiple segments and run each segment on a different machine. **Mahout** is such a data mining framework that normally runs coupled with the Hadoop infrastructure at its background to manage huge volumes of data.

What is Apache Mahout?

A *mahout* is one who drives an elephant as its master. The name comes from its close association with Apache Hadoop which uses an elephant as its logo.

Hadoop is an open-source framework from Apache that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.

Apache **Mahout** is an open source project that is primarily used for creating scalable machine learning algorithms. It implements popular machine learning techniques such as:

- Recommendation
- Classification
- Clustering

Apache Mahout started as a sub-project of Apache's Lucene in 2008. In 2010, Mahout became a top level project of Apache.

Features of Mahout

The primitive features of Apache Mahout are listed below.

- The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment. Mahout uses the Apache Hadoop library to scale effectively in the cloud.

- Mahout offers the coder a ready-to-use framework for doing data mining tasks on large volumes of data.

- Mahout lets applications to analyze large sets of data effectively and in quick time.
- Includes several MapReduce enabled clustering implementations such as k-means, fuzzy k-means, Canopy, Dirichlet, and Mean-Shift.

- Supports Distributed Naive Bayes and Complementary Naive Bayes classification implementations.

- Comes with distributed fitness function capabilities for evolutionary programming.

- Includes matrix and vector libraries.

Applications of Mahout

- Companies such as Adobe, Facebook, LinkedIn, Foursquare, Twitter, and Yahoo use Mahout internally.

- Foursquare helps you in finding out places, food, and entertainment available in a particular area. It uses the recommender engine of Mahout.

- Twitter uses Mahout for user interest modelling.

- Yahoo! uses Mahout for pattern mining.

2. MAHOUT—MACHINE LEARNING

Apache Mahout is a highly scalable machine learning library that enables developers to use optimized algorithms. Mahout implements popular machine learning techniques such as recommendation, classification, and clustering. Therefore, it is prudent to have a brief section on machine learning before we move further.

What is Machine Learning?

Machine learning is a branch of science that deals with programming the systems in such a way that they automatically learn and improve with experience. Here, learning means recognizing and understanding the input data and making wise decisions based on the supplied data.

It is very difficult to cater to all the decisions based on all possible inputs. To tackle this problem, algorithms are developed. These algorithms build knowledge from specific data and past experience with the principles of statistics, probability theory, logic, combinatorial optimization, search, reinforcement learning, and control theory.

The developed algorithms form the basis of various applications such as:

- Vision processing
- Language processing
- Forecasting (e.g., stock market trends)
- Pattern recognition
- Games
- Data mining
- Expert systems
- Robotics

Machine learning is a vast area and it is quite beyond the scope of this tutorial to cover all its features. There are several ways to implement machine learning techniques, however the most commonly used ones are **supervised** and **unsupervised learning**.

Supervised Learning

Supervised learning deals with learning a function from available training data. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. Common examples of supervised learning include:

- classifying e-mails as spam,
- labeling webpages based on their content, and
- voice recognition.

There are many supervised learning algorithms such as neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers. Mahout implements Naive Bayes classifier.

Unsupervised Learning

Unsupervised learning makes sense of unlabeled data without having any predefined dataset for its training. Unsupervised learning is an extremely powerful tool for analyzing available data and look for patterns and trends. It is most commonly used for clustering similar input into logical groups. Common approaches to unsupervised learning include:

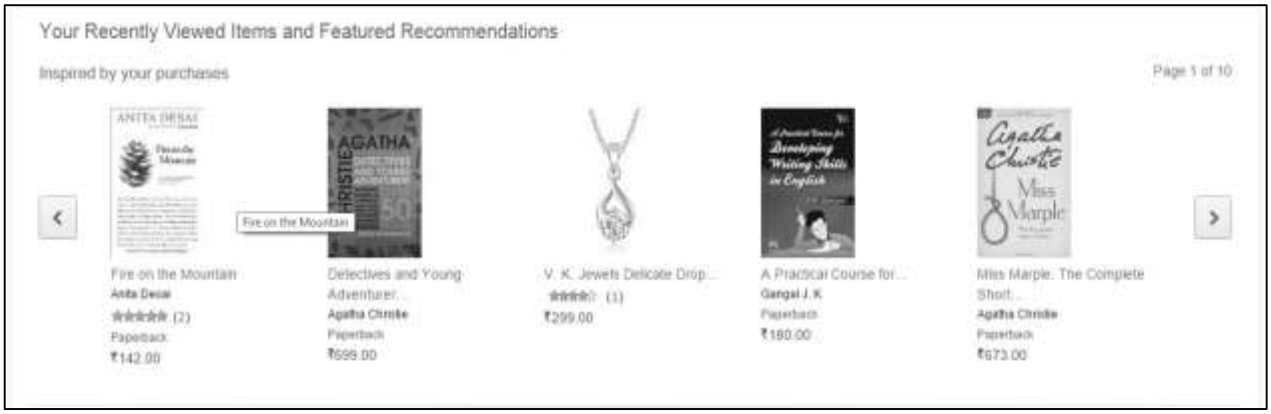
- k-means,
- self-organizing maps, and
- hierarchical clustering.

Recommendation

Recommendation is a popular technique that provides close recommendations based on user information such as previous purchases, clicks, and ratings.

- Amazon uses this technique to display a list of recommended items that you might be interested in, drawing information from your past actions. There are recommender engines that work behind Amazon to capture user behavior and recommend selected items based on your earlier actions.

- Facebook uses the recommender technique to identify and recommend the “people you may know list”.



Your Recently Viewed Items and Featured Recommendations

Inspired by your purchases

Page 1 of 10

Fire on the Mountain
Anita Desai
★★★★★ (2)
Paperback
₹142.00

Detectives and Young Adventurer...
Agatha Christie
Paperback
₹599.00

V. K. Jewels Delicate Drop...
★★★★☆ (1)
₹299.00

A Practical Course for...
Gangal J. K
Paperback
₹180.00

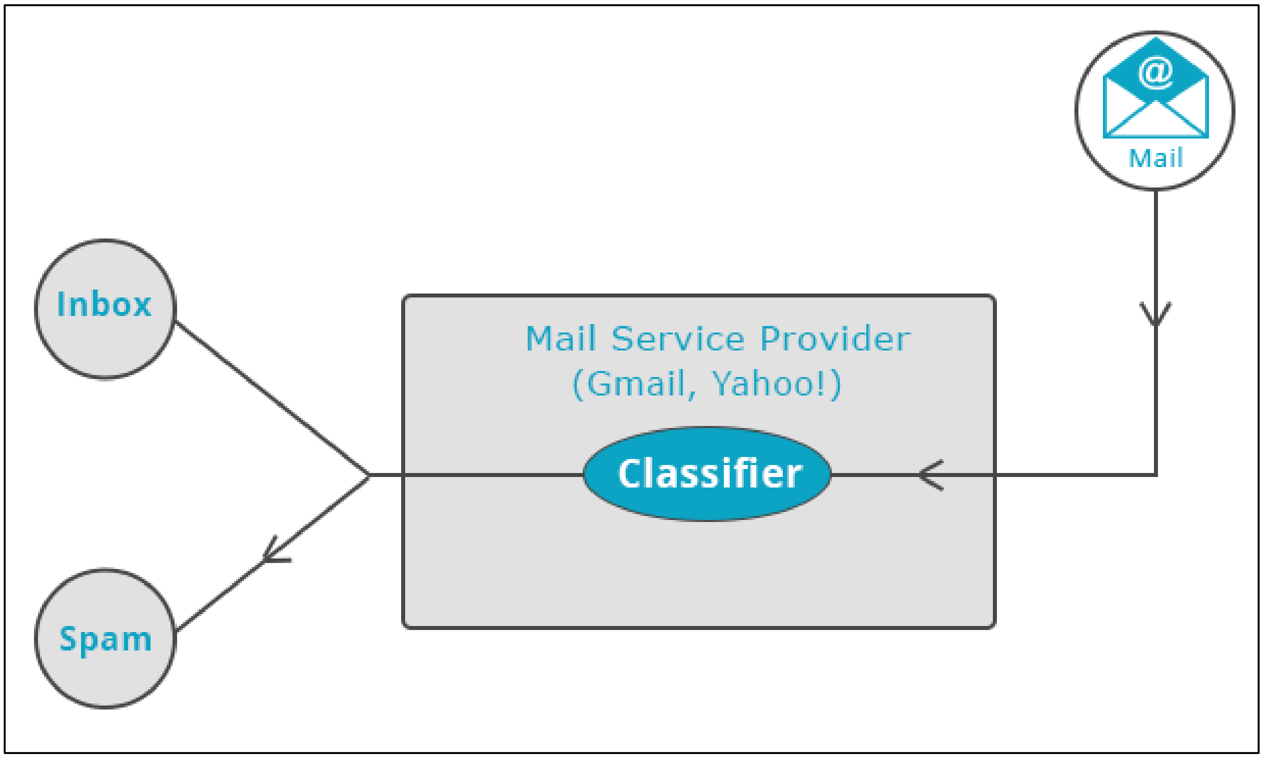
Miss Marple: The Complete Short...
Agatha Christie
Paperback
₹673.00

Classification

Classification, also known as **categorization**, is a machine learning technique that uses known data to determine how the new data should be classified into a set of existing categories. Classification is a form of supervised learning.

- Mail service providers such as Yahoo! and Gmail use this technique to decide whether a new mail should be classified as a spam. The categorization algorithm trains itself by analyzing user habits of marking certain mails as spams. Based on that, the classifier decides whether a future mail should be deposited in your inbox or in the spams folder.

- iTunes application uses classification to prepare playlists.



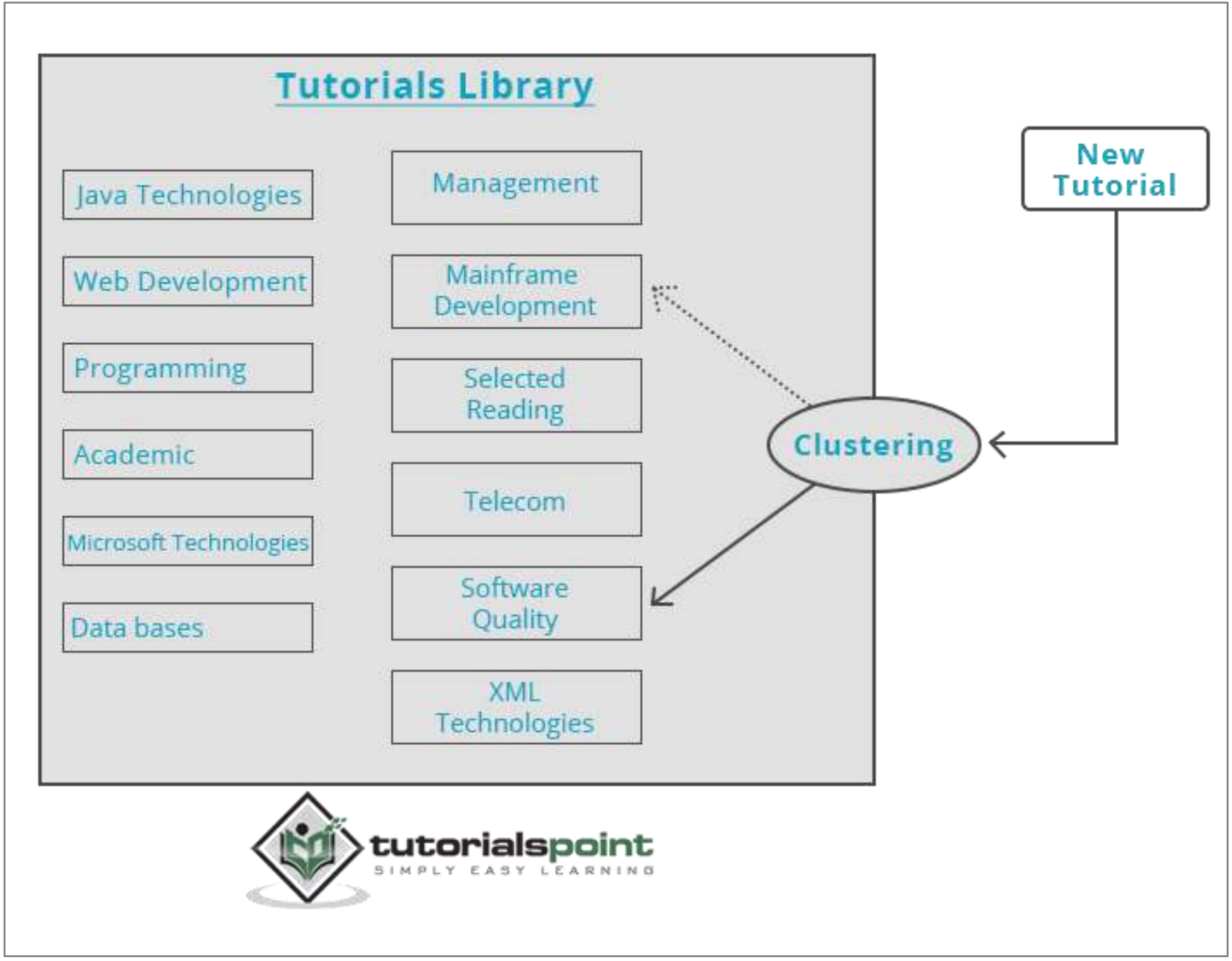
Clustering

Clustering is used to form groups or clusters of similar data based on common characteristics. Clustering is a form of unsupervised learning.

- Search engines such as Google and Yahoo! use clustering techniques to group data with similar characteristics.

- Newsgroups use clustering techniques to group various articles based on related topics.

The clustering engine goes through the input data completely and based on the characteristics of the data, it will decide under which cluster it should be grouped. Take a look at the following example.



Our library of tutorials contains topics on various subjects. When we receive a new tutorial at TutorialsPoint, it gets processed by a clustering engine that decides, based on its content, where it should be grouped.

3. MAHOUT—ENVIRONMENT

This chapter teaches you how to setup mahout. Java and Hadoop are the prerequisites of mahout. Below given are the steps to download and install Java, Hadoop, and Mahout.

Pre-Installation Setup

Before installing Hadoop into Linux environment, we need to set up Linux using **ssh** (Secure Shell). Follow the steps mentioned below for setting up the Linux environment.

Creating a User

It is recommended to create a separate user for Hadoop to isolate the Hadoop file system from the Unix file system. Follow the steps given below to create a user:

- Open root using the command “su”.
- Create a user from the root account using the command “**useradd username**”.
- Now you can open an existing user account using the command “**su username**”.
- Open the Linux terminal and type the following commands to create a user.

```
$ su

password:

# useradd hadoop

# passwd hadoop

New passwd:

Retype new passwd
```

SSH Setup and Key Generation

SSH setup is required to perform different operations on a cluster such as starting, stopping, and distributed daemon shell operations. To authenticate different users of Hadoop, it is required to provide public/private key pair for a Hadoop user and share it with different users.