Module 8: Introduction to Hadoop

Assignment Solution

1. How to get the datasets that come with scikit-learn package?

Solution

from sklearn import datasets

or sklearn.datasets.load_* # datasets is a full package of variety of sample data

2. Give example of two datasets that is shipped with scikit-learn package. Show example how to use one.

Solution

Iris flower dataset which is accessed as “iris” and digits dataset accessed as “digits”

To load it into a variable: digits = datasets.load_digits()

3. List at least 5 modules from scikit-learn?

Solution

sklearn.cluster - Clustering

sklearn.datasets

sklearn.linear_model – Generalized linear models

sklearn.ensemble – Ensemble methods

sklearn.feature_extraction – Feature extraction

4. Take input as few random points in a two dimensional space and divide them into 4 clusters.

5. Use K-means algorithm by using Scikit-learn to find the centroids of each cluster.

6. Plot the points along with the centroids. To distinguish the centroids, use some special symbols like '+' or '#' etc.

Solution (4, 5 & 6)

```
from sklearn.cluster import KMeans

from numpy.random import RandomState

rnge = RandomState(1)

#Initial random data points

x, y = np.random.uniform(0, 10, 100).reshape(2, 25)

# Instantiate model

kmeans = KMeans(n_clusters=4, random_state=rnge)

# Fit model

kmeans.fit(np.transpose((x,y)))

kmeans.cluster_centers_  # To get the centroid coordinates of each cluster

plt.scatter(x, y, c=kmeans.labels_)     # Plots the data points

plt.scatter(*kmeans.cluster_centers_.T, c='r', marker='+', s=100)    # Plots the centroid points

plt.show()
```

7. How many daemon processes run on a Hadoop cluster? Explain.

Solution

Hadoop is comprised of five separate daemons. Each of these daemons runs in its own JVM.

Following 3 Daemons run on Master nodes.NameNode - This daemon stores and maintains the metadata for HDFS.

Secondary NameNode - Performs housekeeping functions for the NameNode.

JobTracker - Manages MapReduce jobs, distributes individual tasks to machines running the Task Tracker.

Following 2 Daemons run on each Slave nodes.

DataNode – Stores actual HDFS data blocks.

TaskTracker – It is Responsible for instantiating and monitoring individual Map and Reduce tasks.

8. If Hadoop spawned 200 tasks for a job and one of the task failed. What will Hadoop do in this case?

Solution

It will restart the task again on some other TaskTracker and only if the task fails more than four (default setting and can be changed) times will it kill the job.

9. Where does mapper output stores the intermediate results?

Solution

The mapper output (intermediate data) is stored on the Local file system (NOT HDFS) of each individual mapper nodes. This is typically a temporary directory location which can be setup in config by the hadoop administrator. The intermediate data is cleaned up after the Hadoop Job completes.

10. When does reducers start their execution in a map reduce process?

Solution

In a MapReduce job reducers do not start executing the reduce method until the all Map jobs have completed. Reducers start copying intermediate key-value pairs from the mappers as soon as they are available. The programmer defined reduce method is called only after all the mappers have finished.