Pyspark based Scalable Machine Learning Algorithm for Handling Soybean Genome Sequences using Biopython

- IIT INDORE

Big Data

- DATA has become an integral part of our life.
- 175 zettabytes of data by 2025 - IDC estimation
- Such an immense quantity of information containing valuable data is called Big Data.

SNP Data

- Extracted from DNA sequences.
- Long sequences of A,T,G,C.
- Goal - find subgroupings in a set of sequences.

Python and Big Data

- Packages
- Easy to use
- Scalability
- Large Community Support
- Compatible with Apache Spark

Apache Spark

- Apache Spark is an open-source cluster-computing framework
- In - memory cache
- RDDs
- ease of use

Pyspark

- **PySpark** is the collaboration of Apache Spark and Python.
- scalable analyses and pipelines
- Lazy execution : In PySpark, operations are delayed until a result is actually needed in the pipeline.

Spark Dataframes and RDDs

- RDD - Resilient Distributed datasets; Fundamental Data structure of spark; Allows to perform in-memory computations
- Dataframes - data organized into named columns; allows developers to impose a structure onto a distributed collection of data allowing higher-level abstraction.

Spark Context

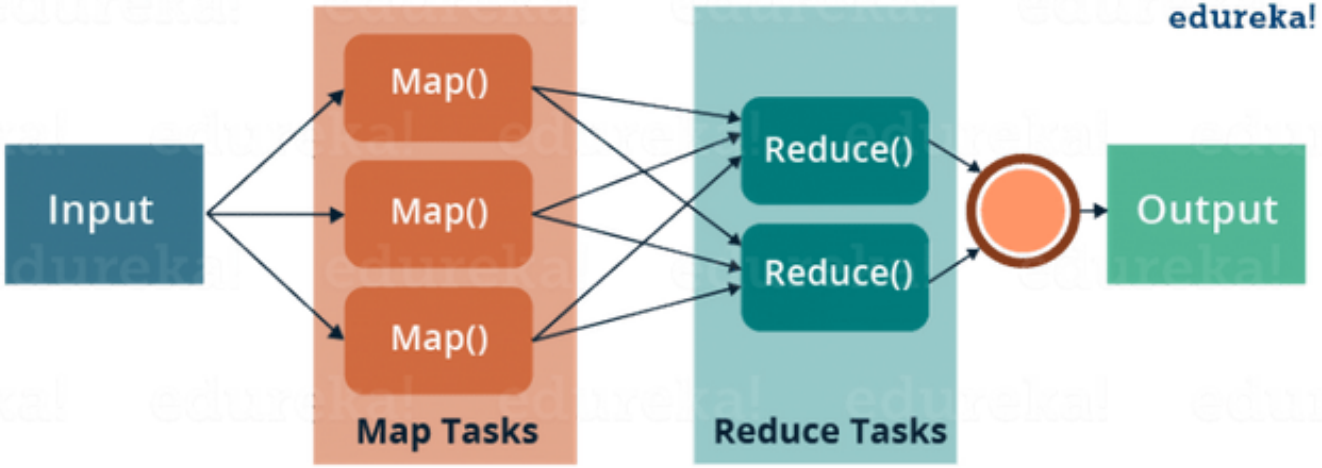
- Spark Context was the entry point of any spark application
- used to access all spark features and needed a sparkConf which had all the cluster configs and parameters to create a Spark Context object.

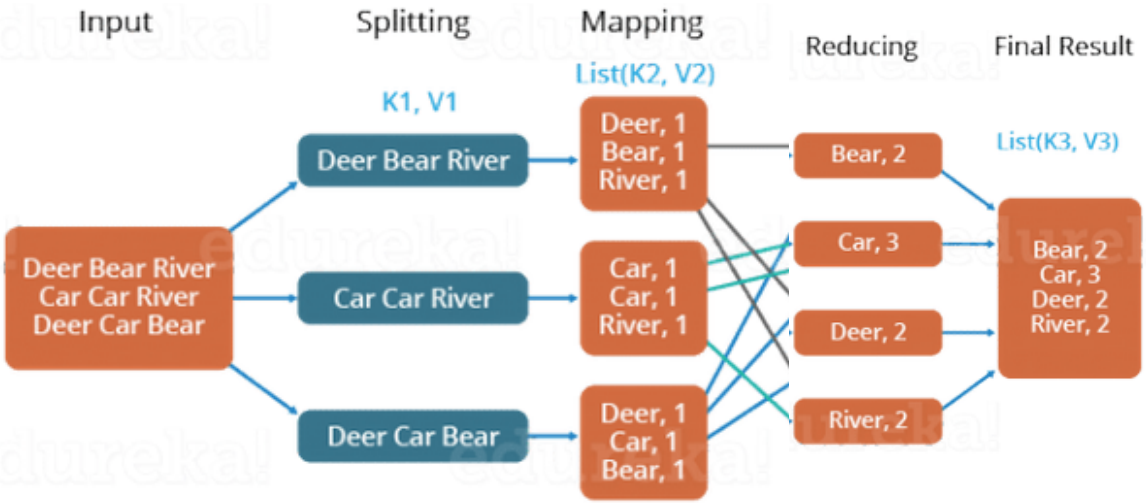
Create RDDs and Dataframes

- RDDs - Resilient Distributed Databases
- Sc = new SparkContext(appName = “Clustering”)
- Rdd = sc.textfile(input_Folder)

Operations on RDD

- Map
  - rdd.map(lambda x : mapper(x))
- Reduce
  - rdd.reduceByKey(lambda x,y : reducer(x,y))


edureka!
Input
Map()
Map()
Map()
Map Tasks
Reduce()
Reduce()
Reduce Tasks
Output

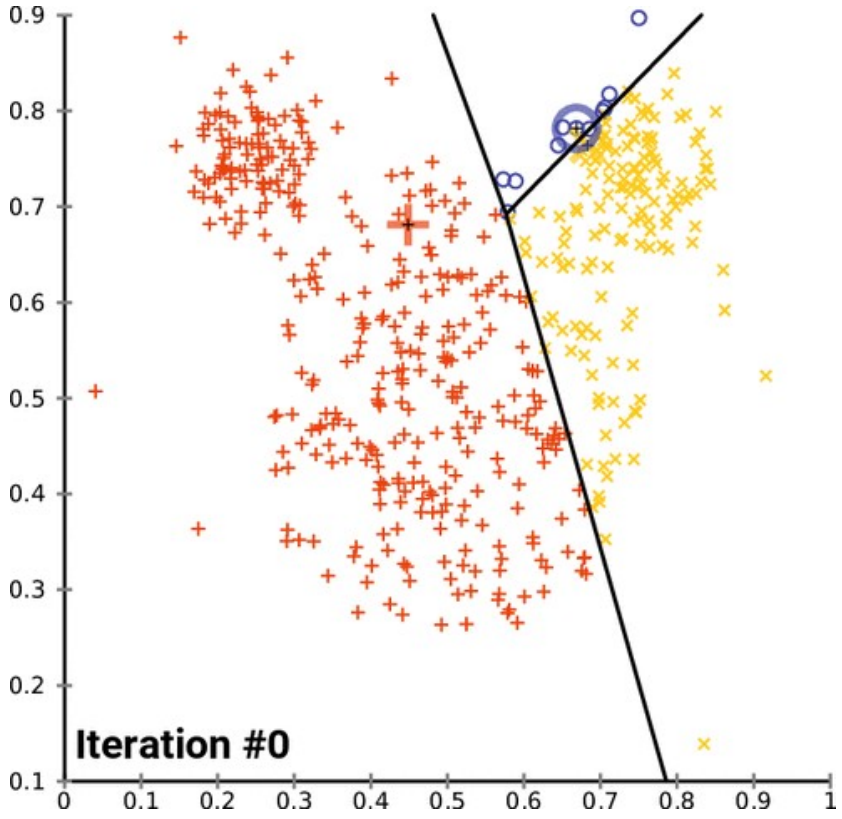

Input
Splitting
Mapping
Reducing
Final Result
K1, V1
List(K2, V2)
List(K3, V3)
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Stop SparkContext

- sc.stop()

IMPLEMENTATION

Fuzzy C-Means Clustering Algorithm



Clustering Algorithms

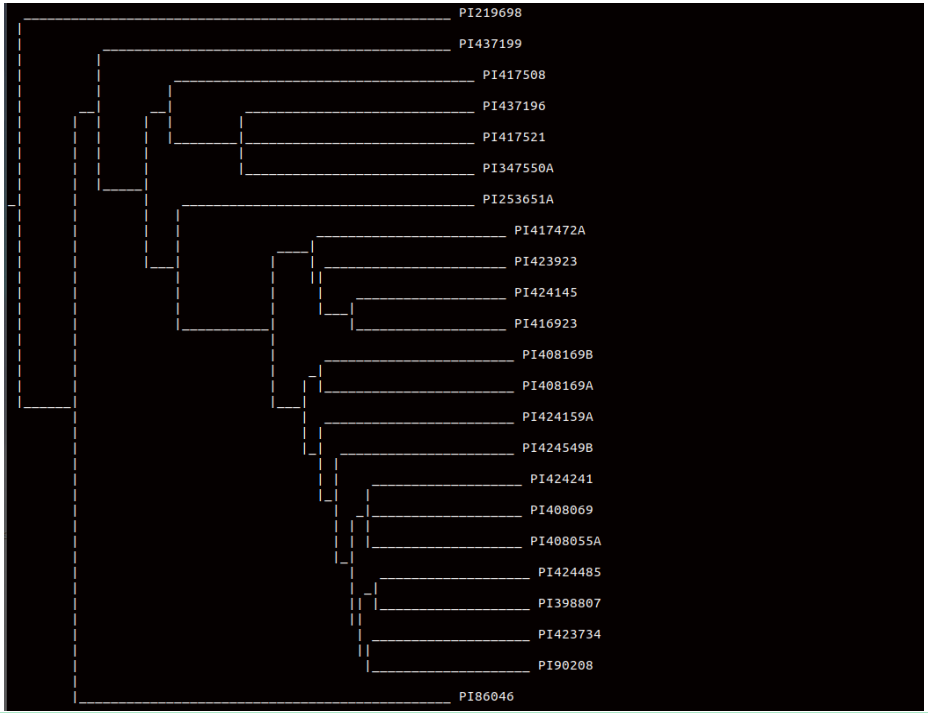
- The results of any clustering algorithm will give us which data point belongs to which cluster.
- Sequences belonging to any cluster exhibit similarities.

What is BioPython?

Image source - https://biopython.org/

- Biopython is a set of freely available tools for biological computation written in Python
- Distributed collaborative effort to develop Python libraries and applications which address the needs of current and future work in bioinformatics.
- https://biopython.org

- Made phylogenetic tree for SNP sequences.
- Unlabelled data
- Used to cross-check Clustering results



THANK YOU!