**Systems of Big Data Assignment Questions**

**You are free to use Hadoop/ Pyspark for solving this assignment.**

**Mandatory:**

1. Try executing the demo exercises for Hadoop/Spark for
2. Matrix Multiplication
3. Word Count

**Codes for the above and process of execution are available in the material shared.**

1. Implement any one Frequent Pattern Mining (FPM) algorithm utilising any larger dataset available over internet like Femi using Hadoop/Pyspark.
2. Test any one classification (or) clustering algorithm utilising any larger dataset available over internet like Femi using Hadoop/Pyspark.

**Practice and Knowledge Purpose**

1. A. What is Commissioning and Decommissioning in terms of Hadoop data nodes. What are the steps involved in each of the process and give practical examples of where these are used.
2. What is the difference between Secondary NameNode and standby NameNode? What does secondary NameNode contain?
3. What are the steps to be followed when a Primary NameNode crashes?
4. A. What are the primary differences between Hadoop 2.x and 3.x architectures?

B. What are the differences between Hadoop and Spark?

C. What are the differences between RDBMS and Hadoop?

D. Write about different data types in Hadoop and differentiate Writable and WritableComparable class

### A. What are the benefits of Spark over MapReduce?

B. What is RDD in Spark? How do you create RDD in Spark?

C. What is Lazy Evaluation in Spark?

D. Write about Key features of Apache Spark?

E. Write about various levels of persistence in Apache Spark?

1. Using the following link

<https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset>

Classify the news by taking equal proportions of data from the available 4 categories of .csv files using advanced classification algorithms.

1. Using the below link related to Portuguese Parliamentary Election dataset , with election result related to 27 parties, cluster the votes in terms of blank, null category available. Use Hadoop ecosystem for clustering.

<http://archive.ics.uci.edu/ml/datasets/Real-time+Election+Results%3A+Portugal+2019>

1. Using the below data set related to Move and TV show rating,

<https://www.kaggle.com/sp1thas/criticker-dataset>

Test drive any two classification algorithms using Hadoop ecosystem

1. Using the following dataset related to Housing price

<https://www.kaggle.com/vaseline555/house-price-of-king-county-extended>

Predict the house prices using regression algorithms

1. Using the MovieLens 25M Dataset from the below link

<https://grouplens.org/datasets/movielens/25m/>

Test various clustering algorithms to cluster movies with respect to rating in Hadoop environment.

1. Using the following Fake News data set

<https://www.kaggle.com/mrisdal/fake-news>

Implement ARBC using Hadoop environment. Also find FI, MFI, CFI for the above dataset.

1. Using the following dataset

<https://www.kaggle.com/c/fake-news/data>

Test various advanced classification algorithms in Hadoop environment to classify news as reliable or unreliable.

1. Using the following link

<https://www.kaggle.com/jacobvs/ddos-attack-network-logs>

Cluster the log files according to various DDOS attack types. Also use classifiers like Random Forest, Naïve Bayes and KNN.

1. Using the following link

<https://www.kaggle.com/jacobvs/ddos-attack-network-logs>

Classify the data with respect to “message “attribute, also cluster the data with respect to “bias” and “audience” attributes.

**Note:** You could use datasets like Femi…

**Some Resources for datasets –** Kaggle, Amazon AWS services, UCI ML repository, US Gov dataset, Data.gov, Analytics Vidhya, Imagenet, MNIST, KDNuggets, MovieLens etc.