Assignment 3
Bigdata Analytics

1. Compare and contrast Resilient Distributed Datasets (RDDs) and DataFrames in Apache Spark. Provide an example where DataFrames offer performance advantages over RDDs. Also, explain how Spark SQL integrates with DataFrames for querying large datasets.
2. Use Spark DataFrames to load a CSV file containing employee information (name, age, salary), and then perform the following operations: filter employees above 30 years of age, group by salary range, and compute the average salary for each group.
3. Using RDDs in Apache Spark, write a program to filter out all the even numbers from a large dataset of integers and compute the sum of all odd numbers. Compare the performance with a traditional single-node approach.
4. Explain how Apache Flink handles real-time stream processing and state management. Discuss the concept of time windows (event time, processing time) and give an example where Flink's stream processing can be used to monitor real-time sensor data for anomaly detection.

Assignment 4
Bigdata Analytics

1. Explain the architecture and working principles of artificial neural networks (ANNs).
2. Discuss the key techniques used in NLP such as tokenization, stemming, and word embeddings. Provide an example of how NLP can be applied to a large dataset for sentiment analysis.
3. Explain the concept of ensemble learning in machine learning. Compare and contrast different ensemble techniques like bagging, boosting, and stacking.
4. Discuss model evaluation techniques, focusing on metrics like accuracy, precision, recall, and the F1 score.
