Tecnológico Nacional de México

Instituto Tecnológico de Tijuana

Departamento de Sistemas y Computación

Ingeniería en Sistemas Computacionales

Semestre:

Febrero - junio 2022

Materia:

Datos masivos

Docente:

JOSE CHRISTIAN ROMERO HERNANDEZ

Unidad:

U1

Título del trabajo:

Práctica Evaluatoria - Unidad 1

alumno:

Anahi Del Carmen Hernandez Pablo 18210486

1)Iniciar sesión en SPARK

***C:\Spark\spark-2.4.8-bin-hadoop2.7\bin\spark-shell***

***import org.apache.spark.sql.SparkSession***

***val spark = SparkSession.builder().getOrCreate()***

Texto

Descripción generada automáticamente

2) Carga el archivo CSV llamada "Netflix\_2011\_2016.csv"

***val data=spark.read.option("Header", "true").option("inferSchema", "true")csv("Netflix\_2011\_2016.csv")***

***val data = spark.read.option("header", "true").option("inferSchema","true")csv("C:/Users/52664/Downloads/Netflix\_2011\_2016.csv")***

Texto

Descripción generada automáticamente

3) ¿Cuáles son los nombres de las columnas?

***data.columns***



4) ¿Cómo es el esquema?

***data.printSchema()***

Texto

Descripción generada automáticamente

5) Imprime las primeras 5 columnas.

***Data.columns.take(5)***

Texto

Descripción generada automáticamente

6) Usa describe () para aprender sobre el DataFrame.

***data.columns.take(5)***

Captura de pantalla de computadora

Descripción generada automáticamente

7. Crea un nuevo dataframe con una columna nueva llamada “HV Ratio” que es la relación que existe entre el precio de la columna “High” frente a la columna “Volumen” de acciones negociadas por un día. Hint - es una operación

***Val df2= df.withColumn(“HV Ratio”, df(“High”)/ df(“volume”))***

***df2.describe().show()***

***Texto

Descripción generada automáticamente***

8. ¿Qué día tuvo el pico mas alto en la columna “Open”?

***df.orderBy(“Open”.desc).show()***

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

9. ¿Cuál es el significado de la columna Cerrar “Close” en el contexto de información financiera, explíquelo no hay que codificar nada?

Es el momento donde se revisa y reducens los balances antes de que el ciclo contable cierre

10. ¿Cuál es el máximo y mínimo de la columna “Volumen”?

***df.select(max(“Volume”)).show()***

***df.select(min(“Volume”)).show()***

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

11. Con Sintaxis Scala/Spark $ conteste los siguiente:

a. ¿Cuántos días fue la columna “Close” inferior a $ 600?

b. ¿Qué porcentaje del tiempo fue la columna “High” mayor que $ 500?

c. ¿Cuál es la correlación de Pearson entre columna “High” y la columna “Volumen”?

d. ¿Cuál es el máximo de la columna “High” por año?

e. ¿Cuál es el promedio de columna “Close” para cada mes del calendario?

Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

***Texto

Descripción generada automáticamente***