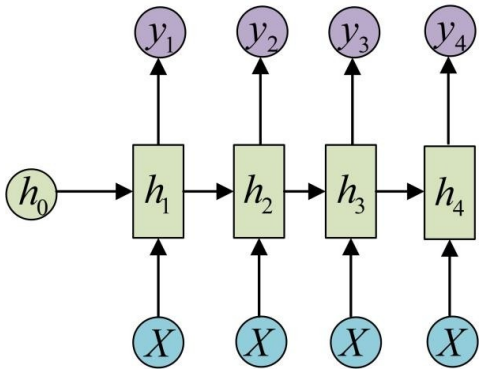
选择题和 UOW 一样

Explain the role of the deep neural network in Big Data Analytics.

Deep neural networks (DNNs) are a type of machine learning model that has proven to be very effective in a wide range of tasks, including big data analytics.
In big data analytics, DNNs are used to automatically learn complex and abstract representations of large and high-dimensional data. These representations can then be used to perform various tasks, such as classification, regression, clustering, or dimensionality reduction.
For example, in image classification, a DNN can be trained on a large dataset of labeled images, and then used to automatically classify new images into different categories, such as dogs, cats, cars, and so on.
Another example is in natural language processing (NLP), where DNNs can be used to analyze large volumes of text data, such as customer reviews, to extract valuable information, such as sentiment analysis or named entity recognition.
Overall, the use of DNNs in big data analytics has enabled new levels of automation and accuracy in many data-driven tasks. However, the training and deployment of DNNs on big data can be computationally intensive and requires specialized hardware and software.
用中文简述一遍
深层神经网络（DNN）是一种机器学习模型，在大数据分析中有很重要的作用。
在大数据分析中，DNN 能够自动学习大量高维数据的复杂抽象表示。这些表示可以用于执行各种任务，如分类、回归、聚类或降维。
例如，在图像分类中，DNN 可以在大量标记图像的数据集上进行训练，然后用于自动将新图像分类为不同的类别，如狗、猫、汽车等。
另一个例子是自然语言处理（NLP），DNN 可以用于分析大量文本数据，例如客户评论，以提取有价值的信息，如情感分析或命名实体识别。
总的来说，DNN 在大数据分析中的使用已经在很多数据驱动的任务中实现了新的自动化和准确性水平。然而，在大数据上训练和部署 DNN 需要大量计算资源，并需要专门的硬件和软件。

Explain one deep neural network, such as CNN, or RNN.



RNN (Recurrent Neural Network) is a special artificial neural network with a "loop" structure that can process sequence data. Its basic unit is the cyclic unit, where the hidden state can capture information at any time step in the sequence. RNNs can be used in many fields such as natural language processing (NLP) and speech recognition.
RNNs consist of three parts: input layer, hidden layer, and output layer. Relative to feedforward neural networks, RNNs can receive hidden states from the previous point in time
The input to the RNN is a sequence with a feature vector for each time step, and the hidden state is connected to the next time step by looping. The hidden state contains information for all previous time steps, so historical information in the sequence can be recorded.

RNNs differ in that their hidden state is cyclical, and the computation of each time step can rely on the previous hidden state. The circular structure of RNNs enables it to learn long-term dependencies, such as the grammatical structure or context of a statement.

However, RNNs have a vanishing/exploding gradient problem, so they are less suitable for long sequence data. To solve this problem, the researchers introduced long short-term memory networks such as LSTM and GRU to better handle long sequence data.

RNN（Recurrent Neural Network）是一种特殊的人工神经网络，具有“循环”的结构，可以处理序列数据。它的基本单元是循环单元，其中的隐藏状态可以捕获序列中任意时间步的信息。
RNN 可以用于许多自然语言处理（NLP）和语音识别等领域。
RNN 包括三个部分：输入层、隐藏层和输出层。相对于前馈神经网络，RNN 可以接收上一个时间点的隐藏状态
RNN 的输入是一个序列，每个时间步输入一个特征向量，隐藏状态通过循环连接到下一个时间步。隐藏状态包含了之前所有时间步的信息，因此可以记录序列中历史信息。

RNN 的不同之处在于其隐藏状态是循环的，每个时间步的计算都可以依赖之前的隐藏状态。
RNN 的循环结构使其能够学习长期依赖关系，例如语句的语法结构或语境。

然而，RNN 存在梯度消失/爆炸问题，因此它们不太适用于长序列数据。为了解决这个问题，研究人员引入了 LSTM 和 GRU 等长短期记忆网络，以更好地处理长序列数据

A dataset describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 27753444 ratings and 1108997 tag applications across 58098 movies. These data were created by 283228 users between January 09, 1995 and September 26, 2018. This dataset was generated on September 26, 2018. The main movie categories are Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western.

Assume you are a movie investor, you would like to summarize some insights into the movie marking by analyzing the MovieLens dataset.

Design a big data analytics project for the above business needs based on the Big Data Analytics Lifecycle in 6 phases: discovery, data preparation, model planning, model building, communicate results, and operationalization. (4 marks) Introduce detail plans of each phase. (6 marks)

Phase 1: Discovery

In this phase, we need to gather the required data from the MovieLens dataset and understand the business needs of the movie investor. We need to identify the key questions that need to be answered, such as:

What are the most popular movie genres among users?

What are the average ratings for different movie genres?

How do the movie ratings and tags relate to each other?

What are the factors that influence the rating of a movie?

We also need to explore the data and understand its structure, quality, and distribution.

Phase 2: Data Preparation

In this phase, we need to clean and prepare the data for analysis. This may involve dealing with missing or inconsistent values, converting the data into a usable format, and transforming the data to better fit the analysis. For example, we might need to group the data by movie genres, and then calculate the average rating for each genre.

Phase 3: Model Planning

In this phase, we need to plan the type of analysis we will perform. For this dataset, we could perform exploratory data analysis (EDA) to understand the distribution of movie ratings and genres, and identify relationships between the data. To answer the business questions, we could also perform regression analysis to identify the factors that influence movie ratings.

Phase 4: Model Building

In this phase, we need to perform the actual analysis using tools such as Python and its libraries such as Pandas and Matplotlib. We can perform regression analysis using linear regression or decision tree regression to identify the factors that influence movie ratings. We can also use clustering algorithms such as k-means or hierarchical clustering to group the movies based on their ratings and tags.

Phase 5: Communicate Results

In this phase, we need to present the results of the analysis in a clear and understandable way. This might involve creating graphs, charts, and tables to visualize the results, and summarizing the key findings in a report or presentation.

Phase 6: Operationalization

In this phase, we need to turn the results of the analysis into actionable insights. For example, based on the results of the analysis, we might recommend the movie investor to focus on certain movie genres that have high potential for success based on the average rating and user interest. We could also use the insights to help inform future movie investments and marketing efforts.

Explain Graph Neural Networks. Also explain why graph modelling is increasingly important to Big Data Analytics.

Graph Neural Networks are a class of deep learning algorithms that operate on graph data. A graph is a data structure composed of nodes (representing entities) and edges (representing the relationships between entities). The aim of GNNs is to learn the representations of graphs and capture the structures and relationships among nodes.

GNNs are becoming increasingly important for big data analysis, as many problems can be represented as graph data. For example, in social network, individuals can be represented as nodes, and their relationships (such as friendship, cooperation, etc.) can be represented as edges. In recommendation systems, items (such as movies or products) can be represented as nodes, and user-item interactions (such as ratings or purchases) can be represented as edges. In these and many other applications, the graph structure of the data provides crucial information about the relationships among entities, which can be used for prediction or decision-making.

GNNs are also highly suitable for processing large and complex graphs because they can operate on graph data in a scalable and distributed manner. This makes them a powerful tool for big data analysis, where the volume, velocity, and variety of data can pose challenges for traditional machine learning methods.

Graph Neural Networks（GNNs）是一类操作图形数据的深度学习算法。图是一种由节点（代表实体）和边（代表实体之间的关系）组成的数据结构。GNN 的目的是学习图的表示，捕捉节点间的结构和关系。

GNN 对大数据分析越来越重要，因为许多实际问题可以被表示为图形数据。例如，在社交网络分析中，个体可以被表示为节点，它们的关系（友谊、合作等）可以被表示为边。在推荐系统中，项目（如电影或产品）可以被表示为节点，用户-项目互动（如评分或购买）可以被表示为边。在这些和许多其他应用中，数据的图形结构提供了关于实体间关系的重要信息，可以用于预测或做出决策。

GNN 还非常适合处理大型复杂图形，因为它们可以以可扩展和分布式的方式操作图形数据。这使得它们成为大数据分析的强大工具，其中数据的量、速度和多样性可能是传统机器学习方法的挑战。

Describe the situation for which you will prefer to use Student's t-test, Welch's t-test, and Wilcoxon Rank-Sum test to conduct hypothesis testing, respectively?

1. Student's t - test（独立样本 t 检验）

适用情况：

当比较两组独立样本的均值是否有显著差异时使用。

假设两个种群的人数相同，但方差未知，并且假设每个种群都是正态分布的（即两组样本需满足正态分布，且方差齐性）。

例如，比较男性和女性的平均身高，假设从男性群体和女性群体中分别随机抽取了两组独立样本，若样本数据满足上述条件，就可以使用 Student's t - test。

2. Welch's t - test（韦尔奇 t 检验）

适用情况：

也是用于比较两组独立样本的均值是否有显著差异。

主要用于人口方差不相等的情况，即当等方差假设不成立时使用（它不要求两组样本方差齐性）。

例如，比较两种不同教学方法下学生的考试成绩，从使用方法 A 的学生群体和使用方法 B 的学生群体中分别抽取样本，如果这两组样本的方差不相等，就应该使用 Welch's t - test。

3. Wilcoxon Rank - Sum test（威尔科克森秩和检验）

适用情况：

用于比较两组独立样本的分布是否有显著差异，它是非参数检验方法。

主要用于两个总体都不是正态分布或者不能假设为正态分布的情况（对数据的分布形态没有严格要求）。

例如，比较两种不同药物对患者康复时间的影响，若康复时间的数据不呈正态分布，就应选用 Wilcoxon Rank - Sum test 来检验两组药物的效果是否有差异。

Explain the following concepts in hypothesis testing: significance level, p-value, t-statistic, and confidence interval.

Significance level: 显著性水平

显著性水平：当原假设为真时，拒绝原假设的概率
the probability of rejecting the original hypothesis, when the null hypothesis is actually TRUE

t - statistics: T 统计
假设两个总体的分布具有相等但未知的方差
假设每个种群都是正态分布的
Suppose that the distributions of two populations have equal but unknown variances
Suppose that each population is normally distributed

p - value: p 值
p 值提供了观察到|T|≥t 的概率，前提是假设为真
p - value offers the probability of observing $|T| \geq t$ given the null hypothesis is TRUE.

Confidence interval: 置信区间
置信区间是指由样本统计量所构造的总体参数的估计区间
Confidence interval is the estimation interval of population parameters constructed by sample statistics
置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度，其给出的是被测量参数的测量值的可信程度，即前面所要求的“一个概率”。

- ARIMA
  - AutoRegressive Integrated Moving Average
  - Shall be applied to stationary time series