Profundiza más

PUCE Virtual

Recurso de Profundización
Clase 5

La **biblioteca** Spark **Mlib** tiene algunos algoritmos de *machine learning*, entre los más importantes **podemos** citar:

1. Clasificación y Regresión

- **Regresión logística**
- **Máquinas de soporte vectorial (SVMs)**
- **Árboles de decisión**
- **Random Forest**
- **Gradient-Boosted Trees (GBT)**
- **Regresión lineal y polinómica**

2. Clustering

- **K-Means**
- **Latent Dirichlet Allocation (LDA)**
- **Gaussian Mixture Model (GMM)**
- **Bisecting K-Means**

3. Reducción de Dimensionalidad

- **PCA (Análisis de Componentes Principales)**
- **SVD (Descomposición en Valores Singulares)**

Fuente: Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2016). Apache Spark. Recuperado de https://en.wikipedia.org/wiki/Apache_Spark

Profundiza más

La librería ***pyspark.sql.functions*** tiene múltiples **funciones**, entre las más importantes podemos **citar**:

1. Funciones Matemáticas

- `abs(col)`: Devuelve el valor absoluto de la columna especificada.
- `sqrt(col)`: Calcula la raíz cuadrada del valor en la columna dada.
- `exp(col)`: Calcula el exponencial del valor en la columna.
- `log(col)`: Devuelve el logaritmo natural del valor de la columna.
- `round(col, scale)`: Redondea el valor de la columna al número de decimales especificado por `scale`.

2. Funciones de Cadenas

- `concat(*cols)`: Concatena múltiples columnas en una sola.
- `substring(col, pos, len)`: Extrae una subcadena de la columna, comenzando en la posición `pos` y con una longitud de `len`.
- `upper(col)`: Convierte todos los caracteres de la columna a mayúsculas.
- `lower(col)`: Convierte todos los caracteres de la columna a minúsculas.
- `trim(col)`: Elimina los espacios en blanco al inicio y al final de la cadena en la columna.

3. Funciones de Fecha y Hora

- `current_date()`: Devuelve la fecha actual.
- `current_timestamp()`: Devuelve la marca de tiempo actual.
- `datediff(end, start)`: Calcula la diferencia en días entre dos fechas.
- `date_add(start, days)`: Suma un número específico de días a una fecha.
- `date_sub(start, days)`: Resta un número específico de días a una fecha.

Fuente: Apache Spark. (s.f.). Spark SQL, Built-in Functions. Recuperado de https://spark.apache.org/docs/latest/api/sql/index.html