x7toktoiw

March 3, 2025

1 RDD's Key-Value

Las funciones Key-Value en PySpark operan sobre RDDs de pares clave-valor, es decir, cada elemento del RDD es una tupla (clave,valor).

Permiten realizar transformaciones y agregaciones eficientes en distribuciones de datos.

[0]:
```
# Datos de ejemplo
rdd = sc.parallelize([(1, 10), (2, 20), (1, 30), (2, 40), (3, 50)])

# 1. reduceByKey(): Aplica una función de reducción (por ejemplo, suma) a los␣
 ↪valores con la misma clave.

result = rdd.reduceByKey(lambda a, b: a + b).collect()
print("reduceByKey result:", result)
```

```
reduceByKey result: [(1, 40), (2, 60), (3, 50)]
```

[0]:
```
# 2. groupByKey(): Agrupa valores por clave y los devuelve como iteradores.

result = rdd.groupByKey().mapValues(list).collect()
print("groupByKey result:", result)
```

```
groupByKey result: [(1, [10, 30]), (2, [20, 40]), (3, [50])]
```

[0]:
```
# 3. sortByKey(): Ordena el RDD por clave

result = rdd.sortByKey().collect()
print("sortByKey result:", result)
```

```
sortByKey result: [(1, 10), (1, 30), (2, 20), (2, 40), (3, 50)]
```

[0]:
```
# 4. keys(): Devuelve solo las claves

result = rdd.keys().collect()
print("keys result:", result)
```

```
keys result: [1, 2, 1, 2, 3]
```

[0]:
```
# 5. values(): Devuelve solo los valores

result = rdd.values().collect()
print("values result:", result)
```

```
values result: [10, 20, 30, 40, 50]
```

[0]:
```
# 6. mapValues(): Aplica una transformación a los valores, manteniendo las␣
 ↪claves

result = rdd.mapValues(lambda x: x * 2).collect()
print("mapValues result:", result)
```

```
mapValues result: [(1, 20), (2, 40), (1, 60), (2, 80), (3, 100)]
```

2 Ejercicios

2.0.1 1. Media de contactos por edad

El fichero contacts.csv tiene una lista de usuarios de una red social, con los siguientes datos (userID, name, age, contacts):

```
0,Will,33,385
1,Jean-Luc,26,2
2,Hugh,55,221
...
```

Obtener el número de contactos promedio para usuarios de cada edad.

2.0.2 2. Temperaturas mínimas por localización

El fichero "1800.csv" contiene las temperaturas del año 1800 (loc,date,type,temp,...)

```
ITE00100554,18000101,TMAX,-75,,,E,
ITE00100554,18000101,TMIN,-148,,,E,
GM000010962,18000101,PRCP,0,,,E,
EZE00100082,18000101,TMAX,-86,,,E,
```

Obtener la temperatura mínima para cada localización.

2.0.3 3. Contar palabras de un libro

Descargar el Quijote de la biblioteca Gutenberg y mostrar las 30 palabras que más se repiten, sin tener en cuenta mayúsculas y minúsculas, y considerando los signos de puntuación como separadores.

Pista: La siguiente expresión convierte el texto a minúsculas, separa las palabras eliminando signos de puntuación y otros caracteres no alfanuméricos y devuelve una lista de palabras lista para análisis de texto (tokenización).

```
re.compile(r'\W+', re.UNICODE).split(text.lower())
```