Ciencia de datos: fundamentos y herramientas (Posgrado)

Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 5

1. **Problema 1: Contador de palabras**
   - Datos a utilizar: archivos de la carpeta "**data/datos_map_reduce/libros**" alojada en https://shorturl.at/pCT78

Dado el archivo "libro1.txt", se requiere contar la cantidad de ocurrencias de cada palabra. La salida debe ser por ejemplo:

...
so 56
what 356
wonderful 3
…

Completar las partes del map y reduce utilizando los scripts mapper.py y reducer.py que se encuentran en el mismo directorio que los datos.

NOTA: por razones de simplificar el ejercicio, no se trabajará en un entorno de Hadoop. Para simular el pipeline del proceso Map-Reduce, la línea de comando a correr es:

```
>> cat libro3.txt | python mapper.py | sort -k1,1 | python reducer.py
```

2. **Problema 2: Matriz de distancia (Opcional)**
   - Datos a utilizar: archivos de la carpeta https://drive.google.com/drive/folders/1UxiHzAG7Et0SSYOUXOslogPfg_jjWeKb?usp=sharing

En el archivo "shortest-path-distance-matrix.txt", leer el encabezado del mismo para informarse de lo que trata dicho archivo.

Ciencia de datos: fundamentos y herramientas (Posgrado)

Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 5

El objetivo de este ejercicio es determinar la cantidad total de cada una de las longitudes existentes entre artículos, es decir, determinar cuál es el total de caminos de cada longitud. Se requiere armar las funciones map y reduce del paradigma Map-Reduce.

2. **Usando Databricks + PySpark:**

a) Generar una cuenta en Databricks Community: https://community.cloud.databricks.com/

b) Dentro de la cuenta generada en el punto anterior:

   a) Crear una instancia cluster.

   b) Generar un espacio de trabajo y crear una notebook.

   c) Reconocer comandos básicos de databricks y PySpark.

c) Datos a utilizar: archivos de la carpetahttps://drive.google.com/drive/folders/1UxiHzAG7Et0SSYOUXOslogPfg_jjWeKb?usp=sharing

- Descagrar los 2 archivos de la carpeta y subirlos a su instancia de Databricks.
- Utilizando los comandos de Databricks, hacer una copia del archivo original (para tener un backup de este).
- Leer por separado ambos archivos de extensión ".csv" que se encuentran en la carpeta mencionada más arriba y almacenarlos en un DataFrame.
- Para cada archivo, analizar: cantidad de columnas, cantidad de datos, tipo de datos.
- Para el archivo de los datos de vuelos, armar un DataFrame que contenga solo las siguientes columnas:
  - `['OP_CARRIER', 'ORIGIN', 'DEST', 'DEP_TIME', 'DEP_DELAY', 'ARR_TIME','ARR_DELAY','DISTANCE', 'FL_DATE']`
- Se desea armar un DataFrame que contenga los datos de los vuelos y además la descripción de cada una de las empresas. (puede usar el método Join de un DataFrame)
- Obtener la cantidad de vuelos de cada una de las empresas. (puede usar el método Count de un DataFrame)

Ciencia de datos: fundamentos y herramientas (Posgrado)

Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 5

- Obtener el tiempo de demora total de cada vuelo. Esto se trata de una columna calculada. (puede usar el método withColumn de un DataFrame y darle a ésta un valor calculado).
- Obtener el promedio de demora de cada una de las empresas de vuelo. (puede usar el método groupBy de un Dataframe combinado con el método Avg).
- Hacer un plot de barras horizontales mostrando el resultado del punto anterior.