UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

DBlandIT®

Spark

Operaciones básicas

1. Crear un rdd "**textRDD**" leyendo el archivo "**/war-and-peace.txt**".
2. Contar la cantidad de elementos en "**textRDD**".
3. Imprimir en pantalla los primeros 10 elementos de "**textRDD**".
4. Imprimir en pantalla la cantidad de palabras en "**textRDD**"

(Opcional)

A. Obtener cuantas veces aparecen dos palabras en una misma linea
B. Obtener cuantas veces aparecen dos palabras consecutivas en la misma línea (bigramas)
C. Idem B, pero 3 palabras consecutivas por línea (trigramas)

Titanic

1. Crear un rdd "**titanicRDD**" leyendo el archivo "**/titanic.csv**"
2. Obtener máximo precio de pasaje por sexo
3. Obtener mínimo precio de pasaje por sexo
4. Obtener cuantos el conteo de cuántos pasajeros masculinos y femeninos sobrevivieron y no sobrevivieron
5. Obtener el promedio de precio por pasaje
    a. ¿Hay alguna diferencia de precio si analizamos por sexo?
    b. ¿Y por clase? (Pclass)

Movielens

1. Crear un rdd "**moviesRDD**" leyendo el archivo "**/ratings.csv**"
2. ¿Cuántos ratings hace en un promedio un usuario?
3. ¿Qué día se hicieron más ratings?

(Opcional)

A. Armar la matriz de co-ocurrencia de ratings de películas (cuántas veces dos películas fueron rateadas por el mismo)