PráCtica sesión 2 Spark -

proyecto 2

El objetivo del siguiente proyecto, es evaluar lo visto en la sesión y en los ejercicios.

**Sumario**

[Ejercicio final Repaso (Proyecto 2) 1](#__RefHeading___Toc631_2103982880)

## Ejercicio final Repaso (Proyecto 2)

Trabajamos en un colegio y tenemos las notas de Matemáticas, Inglés y Física de los alumnos del colegio en 3 documentos txt, a partir de estos ficheros:

*Notas\_Fisica.txt*

*Notas\_Mates.txt*

*Notas\_Ingles.txt*

a) Crea 3 RDD de pares, uno para cada asignatura, con los alumnos y sus notas

|  |  |
| --- | --- |
| SOL | notas\_ingles = sc.textFile('Notas\_Ingles.txt')  pairs\_ingles= notas\_ingles.map(lambda x: (x.split(',')[0], float(x.split(',')[1])))  notas\_mates = sc.textFile('Notas\_Mates.txt')  pairs\_mates= notas\_mates.map(lambda x: (x.split(',')[0], float(x.split(',')[1])))  notas\_fisica = sc.textFile('Notas\_Fisica.txt')  pairs\_fisica= notas\_fisica.map(lambda x: (x.split(',')[0], float(x.split(',')[1]))) |

b) Crea un solo RDD con todas las notas

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo = pairs\_ingles.union(pairs\_fisica).union(pairs\_mates) |

c) ¿Cuál es la nota más baja que ha tenido cada alumno?

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo.reduceByKey(lambda x,y:min(x,y)).collect() |

d) ¿Cuál es la nota media de cada alumno?

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo.reduceByKey(add).map(lambda x: (x[0],x[1]/3)).collect() |

e) ¿En qué asignatura suspende más gente?

|  |  |
| --- | --- |
| SOLUCIÓN | pairs\_ingles.filter(lambda x: x[1]<5).count()  pairs\_mates.filter(lambda x: x[1]<5).count()  pairs\_fisica.filter(lambda x: x[1]<5).count() |

f) Total de notables o sobresalientes por alumno

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo.filter(lambda x: x[1]>=7).countByKey() |

g) ¿Qué alumno no se ha presentado a inglés?

|  |  |
| --- | --- |
| SOLUCIÓN | pairs\_mates.subtractByKey(pairs\_ingles).collect() |

h) ¿A cuántas asignaturas se ha presentado cada alumno?

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo.countByKey() |

i) Obtén un RDD con cada alumno con sus notas

|  |  |
| --- | --- |
| SOLUCIÓN | rdd\_completo.groupByKey().collect() |