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)

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

SOLUCIÓN	rdd_Fisica = sc.textFile("Notas_Fisica.txt").map(lambda x: x.split(',')).map(lambda y: (y[0],float(y[1]))) rdd_Ingles = sc.textFile("Notas_Ingles.txt").map(lambda x: x.split(',')).map(lambda y: (y[0],float(y[1]))) rdd_Mates = sc.textFile("Notas_Mates.txt").map(lambda x: x.split(',')).map(lambda y: (y[0],float(y[1])))

b) Crea un solo RDD con todas las notas

SOLUCIÓN	rdd_Total = rdd_Fisica.union(rdd_Ingles).union(rdd_Mates)

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

SOLUCIÓN	rdd_Total.reduceByKey(lambda x,y: min(x,y)).collect()

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

SOLUCIÓN	# La primera no és moltbona, perquè sempre divideix per 3. La segona sí que és bona
	#rdd_mitj = rdd_Total.reduceByKey(lambda x,y: x+y).map(lambda y: (y[0],y[1]/3))
	rdd_Total.groupByKey().mapValues(lambda x: sum(x)/len(x)).collect()

e) ¿En qué asignatura suspende más gente?

SOLUCIÓN	# El més fàcil seria comprovar separadament cada rdd. He intentat muntar un altre rdd que ens puga donar directament el resultat
	rdd_NotesAssig = sc.parallelize([\
	('Fisica',rdd_Fisica.filter(lambda x: x[1]<5.0).count()),\
	('Ingles',rdd_Ingles.filter(lambda x: x[1]<5.0).count()),\
	('Mates',rdd_Mates.filter(lambda x: x[1]<5.0).count())])
	rdd_NotesAssig.takeOrdered(1, key = lambda x: -x[1])

f) Total de notables o sobresalientes por alumno

SOLUCIÓN	# Ens ho dóna en forma de diccionari, però crec que és vàlid
	rdd_Total.filter(lambda x: x[1]>=7).countByKey()

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

SOLUCIÓN	rdd_Total.subtractByKey(rdd_Ingles).keys().distinct().collect()

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

SOLUCIÓN	rdd_Total.groupByKey().map(lambda x: (x[0],len(x[1]))).collect() # ens apareix Rocio que s'ha presentat a 4 perquè apareix 2 vegades en Fisica

i) Obtén un RDD con cada alumno con sus notas

SOLUCIÓN	rdd_notesAlumnes = rdd_Total.groupByKey() # bucle per a poder fer list de les notes [(x,list(y)) for x,y in rdd_notesAlumnes.collect()]