PráCtica sesión 1 Spark -

proyecto 1

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

**Sumario**

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

## Ejercicio final Repaso (Proyecto 1)

A partir de la lista siguiente:

['Alicante','Elche','Valencia','Madrid','Barcelona','Bilbao','Sevilla']

**llistaCiutats = ['Alicante','Elche','Valencia','Madrid','Barcelona','Bilbao','Sevilla']**

**rdd\_llista = sc.parallelize(llistaCiutats)**

a) Usando lo que hemos aprendido hasta ahora, quédate sólo con las ciudades que tengan la letra 'e' en su nombre y muestralas

**rdd\_res1 =rdd\_llista.filter(lambda x: x.find('e')>=0)**

**rdd\_res1.collect()**

b) Muestra las ciudades que tienen la letra ‘e’ y muestra el número de veces que aparece en cada nombre. Por ejemplo (Elche,2)

**rdd\_res2 = rdd\_res1.map(lambda x: (x,x.lower().count('e')))**

**rdd\_res2.collect()**

c) Podrías quedarte solo con las ciudades que tengan una sola e?

**# Imagine que s’està referint de les del resultat anterior.**

**# Sinó, partiríem del rdd incial, i filtraríem per == 1**

**rdd\_res31 = rdd\_res2.filter(lambda x: x[1]==1)**

**rdd\_res31.collect()**

d) Han pasado una nueva lista pero no han separado bien las ciudades.. podrias volver a contar cuantas e hay en cada ciudad?

ciudades\_mal = [['Alicante.Elche','Valencia','Madrid.Barcelona','Bilbao.Sevilla'],['Murcia','San Sebastián','Melilla.Merida']]

**rdd\_llistaMal = sc.parallelize(ciudades\_mal)**

**# Primer flatMap per a passar de matriu a llista, després**

**# flatMap per a separar els elements que hi ha més d’un**

**# Per últim map per a posar el número de e**

**# Si vulguérem traure només les de 1 e, faríem un filter com en c)**

**rdd\_res4 = rdd\_llistaMal.flatMap(lambda x: x).flatMap(lambda y: y.split('.')).map(lambda z: (z,z.lower().count('e')))**

**rdd\_res4.collect()**