Componente Apache Spark.

**En esta práctica se trabajará con la herramienta Apache Spark.**

**Se pide:**

* Investiga el alogritmo de Euclides, explica para que puede ser usado en diferentes situaciones e impleméntalo en código de tal manera que desde Apache Spark se pueda ejecutar de manera distribuida. Describe un ejemplo de ejecución del algoritmo en notación matemática (**Euclid**).
* Teniendo como entrada el fichero ratings.txt en el que cada línea es el código de una película, el código de un usuario, la calificación asignada y el momento de la votación. Crear dos scripts en Python que se puedan ejecutar en Apache Spark de manera distribuida para obtener :
  + Para cada película, la nota media de todas sus votaciones (**movies\_each\_average**).
  + Películas cuya nota media sea superior a 3 (**average\_movie\_mark**).
* Reliza el ejercicio que se hizo en la práctica de MapReduce utilizando el concepto de Dataframe de Apache Spark (**mapreduce**).
* Crea un RDD con la tabla películas de la base de datos de la práctica anterior y muestra las películas que tengan 4 o más vocales en su nombre (**movies**).
* Crea un RDD con todas las palabras que aparezcan en El Quijote y (**quijote**):
  + Crea una lista con todas las palabras del documento.
  + Muestra cuantas veces aparece una palabra que tu elijas (independientemente de si está en mayúsculas o minúsculas). Especifica que palabra es.
  + Devuelve una lista ordenada según el número de veces que sale cada palabra de más a menos. Guarda este resultado en HDFS.
* Dada una lista de elementos desordenados y algunos repetidos, devolver una muestra de 5 elementos, que estén en la lista, sin repetir y ordenados descendentemente (**numbers**). List = 4,6,34,7,9,2,3,4,4,21,4,6,8,9,7,8,5,4,3,22,34,56,98
  + Selecciona el elemento mayor de la lista resultante.
  + Muestra los dos elementos menores.
* Una vez creado un RDD con tuplas de palabras y su traducción de las listas (**words**):

- Inglés: hello, table, angel, cat, dog, animal, chocolate, dark, doctor, hospital, computer.

- Español: hola, mesa, angel, gato, perro, animal, chocolate, oscuro, doctor, hospital, ordenador.

* + Averigua palabras que se escriben igual en inglés y en español.
  + Averigua palabras que en español son distintas que en inglés.
  + Obtén una única lista con las palabras en ambos idiomas que son distintas entre ellas (['hello', 'hola', 'table', ...).
  + Haz dos grupos con todas las palabras, uno con las que empiezan por vocal y otro con las que empiecen por consonante.
* Dada una cadena que contiene una lista de nombres Juan, Jimena, Luis, Cristian, Laura, Lorena, Cristina, Jacobo, Jorge, una vez transformada la cadena en una lista y luego en un RDD (**names**):
  + Agrúpalos según su inicial, de manera que tengamos tuplas formadas por la letra inicial y todos los nombres que comienzan por dicha letra:

[('J', ['Juan', 'Jimena', 'Jacobo', 'Jorge']),

('L', ['Luis', 'Laura', 'Lorena']),

('C', ['Cristian', 'Cristina'])]

* + De la lista original, obtén una muestra de 5 elementos sin repetir valores.
  + Devuelve una muestra de datos de aproximadamente la mitad de registros que la lista original con datos que pudieran llegar a repetirse.
* Usando los ficheros notas\_matematicas, notas\_ingles y notas\_fisica del material dado realizar lo siguiente (**marks**):
  + Crea 3 RDD de pares, uno para cada asignatura, con los alumnos y sus notas
  + Crea un solo RDD con todas las notas
  + ¿Cuál es la nota más baja que ha tenido cada alumno?
  + ¿Cuál es la nota media de cada alumno?
  + ¿Cuantos estudiantes suspende cada asignatura? [('Matematicas', 7), ('Física', 8), ('Inglés', 7)]
  + ¿En qué asignatura suspende más gente?
  + Total de notables o sobresalientes por alumno, es decir, cantidad de notas superiores o igual a 7.
  + ¿Qué alumno no se ha presentado a inglés?
  + ¿A cuántas asignaturas se ha presentado cada alumno?
  + Obtén un RDD con cada alumno con sus notas.

**Material de apoyo:**

* <https://aitor-medrano.github.io/iabd/spark/spark.html>

**A entregar:**

* Los scripts correspondientes de Apache Spark con el nombre que está entre parentesis en negrita en cada apartado.
* Un fichero de TEXTO PLANO con los comandos a ejecutar los scripts de Spark.
* Todo lo anterior se debe de comprimir en un fichero (.zip) llamado IAB\_BIU\_Tema\_03\_Practica\_05\_01\_Nombre\_Apellidos (sin tildes, ni ñ ni espacios). Si el grupo lo forma más de una persona poner:
  + IAB\_BIU\_Tema\_03\_Practica\_05\_01\_Nombre1\_Apellidos1\_Nombre2\_Apellidos2