FACULTAD DE INGENIERÍA Y ARQUITECTURA
ASIGNATURA: Analítica con Big Data
PERIODO ACADÉMICO: 2017-2
FECHA : semana del 2 de Octubre
TIEMPO: -

EXAMEN PARCIAL – PARTE PRÁCTICA

CÓDIGO	APELLIDOS Y NOMBRES	SECCIÓN

INSTRUCCIONES GENERALES:

- La prueba consta de "2" preguntas, cuyo puntaje está indicado en cada una de ellas.
- El examen se debe de realizar en grupos según lo enviado al profesor.
- El procedimiento, el orden, la claridad de las respuestas y el uso apropiado del lenguaje (notaciones, símbolos y unidades), serán considerados como criterios de calificación.
- Puede utilizar código fuente externo pero debe ser referenciado en un comentario poniendo el origen del código.
- Se presentará el trabajo en clase el día Viernes 6 de Octubre además de enviar el código fuente por el blackboard (menu Evaluaciones).
- **Máximo puntaje: 10 puntos.**
- **Leer detenidamente las situaciones que ocasionarán la anulación de la prueba, que se encuentran a continuación.**

SITUACIONES QUE OCASIONARÁN LA ANULACIÓN DE LA PRUEBA:

- Mantener prendidos teléfonos celulares, relojes smart, así como cualquier otro medio o dispositivo electrónico de comunicación.
- Detectarse plagio.
- Utilizar material de consulta no autorizado (apuntes de clase, fotocopias o materiales similares).

Los profesores de la asignatura

CASO Analítica de RRHH

Uno de los algoritmos de clusterización más utilizados es el k-means. Este algoritmo se encarga de poder clasificar elementos de un dataset en base a alguna similitud.

Además, se le está proporcionando un dataset ficticio con datos referentes al desempeño en una empresa XYZ. Los datos que nos proporciona el dataset son los siguientes:

satisfaction_level	Nivel de satisfacción del empleado con la empresa (entre 0 y 1).
last_evaluation	Tiempo en años desde su última evaluación.
number_project	Número de proyectos completados durante su trabajo.
average_monthly_hours	Número de horas promedio trabajadas mensuales.
time_spend_company	Número de años que el empleado está/estuvo en la empresa.
work_accident	Flag que indica si el empleado tuvo un accidente en el lugar de trabajo.
left	Flag que indica si el empleado dejo de trabajar en la empresa.
promotion_last_5years	Flag que indica si el empleado tuvo una promoción en los últimos 5 años.
sales	Área en la que trabaja la persona.
salary	Nivel relativo del salario.

PREGUNTA 1 (5 puntos)

Implementar una función en Scala que nos permita agrupar (clusterizar) un dataset de empleados según un feature que se le proporcione. Esta clusterización deberá realizarse utilizando el algoritmo K-means estándar así como mapReduce como modelo de programación. La cantidad k de centroides (grupos) deberá ser ingresada como parámetros así como el índice del feature a calcular.

Por ejemplo, quiero agrupar según satisfaction_level, entonces deberé pasar a la función los parámetros k=3 y el índice 0. El resultado deberá ser un RDD que tenga todos los features más el centroide (promedio) al que pertenece.

PREGUNTA 2 (5 puntos)

Implementar una función en Scala que nos permita agrupar (clusterizar) un dataset de empleados según **todos** sus features. Debe utilizar el algoritmo K-means estándar tomando como función de costo a minimizar la distancia cuadrada euclidiana de los vectores (https://en.wikipedia.org/wiki/Euclidean_distance). Tomar en cuenta que como en la pregunta anterior, debe tomar como parámetro la cantidad k de centroides, dar como respuesta un RDD de la misma forma que la pregunta 1 y utilizar todos los features. Los features que son alfanuméricos puede normalizarlos utilizando un número.

RÚBRICA POR PREGUNTA

Concepto	Descripción
Funcionalidad	Cumple con todo los pedido (3 puntos máximo)
Comentario	Esta comentado y explicado correctamente todo su código (1)
Estructura	Se encuentra modularizado según lo que se le pide (0.5 puntos)
Presentación	Presenta los resultados en un formato estándar y entendible (archivo). (0.5 puntos)

REFERENCIAS

- https://stanford.edu/~rezab/classes/cme323/S16/projects_reports/bodoia.pdf