**Instituto Tecnológico de Tijuana**

**Ingeniería en Sistemas Computacionales** 

**Examen III**

K-Means

**Materia:** Datos Masivos

**Unidad:** Unidad III

**Facilitador:**

José Christian Romero Sánchez

**Alumnos:**

* Hernández Negrete Juan Carlos - **16212021**
* Manzano Guzman Jesua - **16212033**

**Fecha:**

Tijuana Baja California a 18 de Diciembre del 2020.

**Introduction.**

For this evaluative practice, a K-means clustering method will be developed, Group Analysis or grouping is the task of grouping objects by similarity, in groups or sets so that members of the same group have similar characteristics. It is the main task of exploratory data mining and is a common technique in statistical data analysis.

**Development.**

**EVALUATION PRACTICE UNIT 3**

Import a simple Spark session.

|  |
| --- |
| **import org.apache.spark.sql.SparkSession** |

Use lines of code to minimize errors

|  |
| --- |
| import org.apache.log4j.\_ Logger.getLogger("org").setLevel(Level.ERROR) |

Create an instance of the Spark session

|  |
| --- |
| val spark = SparkSession.builder().getOrCreate() |

Import the K-means library for the clustering algorithm.

|  |
| --- |
| import org.apache.spark.ml.clustering.KMeans |

Loads the Wholesale Customers Data dataset

|  |
| --- |
| val dataset = spark.read.option("header","true").option("inferSchema","true").csv("Wholesale\_customers\_data.csv") dataset.show |

Select the following columns: Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicassen and call this set feature\_data

|  |
| --- |
| val feature\_data = dataset.select($"Fresh", $"Milk", $"Grocery", $"Frozen", $"Detergents\_Paper", $"Delicassen") feature\_data.show |

Import Vector Assembler and Vector

|  |
| --- |
| import org.apache.spark.ml.feature.VectorAssembler |

Create a new Vector Assembler object for the feature columns as an input set, remembering that there are no labels

|  |
| --- |
| val assembler = (new VectorAssembler()  .setInputCols(Array("Fresh","Milk", "Grocery","Frozen","Detergents\_Paper","Delicassen"))  .setOutputCol("features")) |

Use the assembler object to transform feature\_data

|  |
| --- |
| val features = assembler.transform(feature\_data) features.show |

Create a Kmeans model with K = 3

|  |
| --- |
| val kmeans = new KMeans().setK(3).setSeed(1L) val model = kmeans.fit(features) |

Evaluate the clusters using Within Set Sum of Squared Errors WSSSE and print the centroids.

|  |
| --- |
| val WSSSE = model.computeCost(features) println(s"Within Set Sum of Squared Errors = $WSSSE")  println("Cluster Centers: ") model.clusterCenters.foreach(println) |

**Conclusion**

The tools and techniques learned throughout the Big Data course were implemented in order to carry out this evaluative practice in addition to the theoretical knowledge on the subject (K-means), with the problem raised, the clustering method was developed for a set of specific data and because the practice was fully completed, it can be said that the knowledge acquired is put into practice.

**Repository**

[**https://github.com/JesuaMG/BigData/tree/Unit\_3/Unit3/Evaluation**](https://github.com/JesuaMG/BigData/tree/Unit_3/Unit3/Evaluation)

**Video**