UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO

**Roteiro de Aula Prática – Criando Aplicações com Apache Spark em RDD**

DISCIPLINA: DCA0132 – Engenharia de Dados

PROFESSOR: Carlos Manuel Dias Viegas

Esta prática consiste em criar aplicações com Apache Spark programando com RDD na linguagem Python.

- Os arquivos com o código fonte (.py) devem ser submetidos em uma tarefa específica no SIGAA.

- Esta prática pode ser realizada em duplas.

**TAREFAS:**

* Desenvolver uma aplicação sobre o conjunto de dados “conjunto2”, que representa uma folha de pagamento de funcionários de uma empresa qualquer, na qual devem ser lidos todos os cargos e apresentada a quantidade de pessoas nos mesmos. Devem ser utilizados os conceitos de RDD em pySpark. É importante retirar as duplicatas antes de fazer a contagem de pessoas.

Observação: não utilizar Dataframes.

Conjunto de dados disponível em: <https://goo.gl/A3MhFS>

Datasets > conjunto2.csv