Cloud Computing for Data Analysis Group Project Description:

This is a Group Project. Locate your Group Members on Canvas.*

1. Prepare 14 PowerPoint slides (total for the Group) on your assigned subject (as shown below).
2. Create a Demo Video with audio on your implementation - This video should explain all the steps followed in exuting the program and obtaining the results.
3. Implement your assigned algorithm (as shown below). Use - JAVA (or Scala) - as a programming language . Create a User Interface .
4. One student should run a Demonstration of the code before the class , and explain what is the purpose of the code , what inputs it takes , what outputs it produces .
5. Submit your project files: PowerPoint file, Demo video (.mp4 format), and SourceCode to Canvas - due: 3 days prior to your assigned Group Moderator Date presentation date (shown on the syllabus).
6. Present your PowerPoints, Video, and Implementation Demo to the class on your assigned Group Moderator Date. (shown on the syllabus). Presentation should not take more than 15 minutes altogether .
7. Each student presents 2 PowerPoint slide, and speaks for 1 to 2 minutes maximum , and prepare 1 question for the audience based on his / her PowerPoint slides.
8. Answer questions. Each group will ask the Presenters 1 question.
9. Bring CANDY / Sweets (ex. chocolates (sneakers , mars , M&M's , etc.) , cookies , cupcakes , doughnuts) for the audience. Each group will give you a score from 0 to 10 for your Presentation.

10. For ONLINE class Project submissions :::
For GroupProject PRESENTATION : Presenting Group - Project Files are due - 3 days *before* your assigned Presentation Date.
Presenting Group - you are Required to SUBMIT 3 VIDEOS :
_1.1. Video#1: Record YourSelf - talking about your PowerPoint Slides.
Including ASKING 1 QUESTION at the end of your Slides.
After that - PROVIDE THE ANSWER to your question, and record that.
Video#1 should include 1/2 of the GroupMembers recorded.
Example: if there are 8 people in the Group - then Video#1 should include the first 4 people: persons 1, 2, 3, 4
Altogether the Video#1 should NOT BE LONGER than 12 minutes.
Name the video file with the SUBJECT that it covers, for example:
Group01_MapReduceTypes_Formats_Features_Video01.mp4
_1.2. Video#2: - same as Video#1 - should include the SECOND 1/2 of the GroupMembers recorded. Example: if there are 10 people in the Group - then Video#2 should include people: 5, 6, 7, 8.
Altogether the Video#2 should NOT BE LONGER than 12 minutes.
Name the video file with the SUBJECT that it covers, for example:
Group01_MapReduceTypes_Formats_Features_Video02.mp4
_1.3. Video#3: create a DEMO Video of your Code / Programming Assignment Exercise - to explain :
*What is the purpose of this code *What is the Input Data File - show the Data - and explain what it means
*How do we run/execute this code *What is the Output Data file produced - show the Output - and explain what it means
*Show how to remove the AWS cluster - after usage
*Make sure the DEMO Video has SOUND - your VOICE is recorded
*Name the video file with the SUBJECT that it covers, for example:
Group01_02_Exercise_ExampleMapReduceProgram_DEMO.mp4

Group 1

Presentation Subject: MapReduce Types , Formats , and Features
Chapter 8. from Book 2. HadoopTheDefinitiveGuide
Chapter 9. from Book 2. HadoopTheDefinitiveGuide
implement: Run the Example MapReduce Program as described in :
1.
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/02_Exercise_ExampleMapReduceProgram.txt
2.

http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/ExampleMapReduce_ModifiedInstructions.docx
3. http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/InstructionsForDSBAHadoopCluster.txt
4.
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/02_ExampleMapReduceProgram_WithoutCloudera.txt
5.
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/02_ExampleMapReduceProgram_UsingAWS.txt

Group 2

Presentation Subject: Pig | Hive | HBase | Zookeeper
Chapter 16. from Book 2. HadoopTheDefinitiveGuide
Chapter 17. from Book 2. HadoopTheDefinitiveGuide
Chapter 20. from Book 2. HadoopTheDefinitiveGuide
implement: HIVE program as described in :
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/03_Exercise_Hive.txt

Group 3

Presentation Subject: Downloading Spark , Getting Started , Simple Spark Applications , Scala and Python Example Programs | Intro to Scala
Chapter 5. from Book 3. LearningSpark
Chapter 9. from Book 3. LearningSpark
Chapter 11. from Book 3. LearningSpark
implement: Spark SQL program as described in :
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/04_Exercise_SparkSQL.txt

Group 4

Presentation Subject: Boolean Retrieval | Term Vocabulary and Posting Lists | Web Search Basics
Chapter 1. from Book 4. InformationRetrieval
Chapter 2. from Book 4. InformationRetrieval
Chapter 19. from Book 4. Information Retrieval
implement: PageRank program as described in :
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/05_Exercise_PageRank.txt

Group 5

Presentation Subject: Frequent Itemsets , Market Basket , Association Rules , Apriori , Other Algorithms
Read Chapter 6. from Book 1. MiningOfMassiveDatasets
implement: AssociationRulesMining as described in :
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/AgrawalExample.doc
//write a program , which implements the algorithm from this exercise . use the same data from the exercise as an input, and check your output to match the results of the exercise . user should be asked to provide minimum support treshold before program starts .

http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/07_Exercise_Part2_SparkAssociationRules_AWS.txt

Group 6

Presentation Subject: Recommender Systems 01 , Content Based , Collaborative Filtering
Chapter 9. from Book 1. MiningOfMassiveDatasets
implement: DecisionTreeSystemID3 as described in :
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/ID3Example.doc
Example Code:
http://webpages.uncc.edu/aatzache/ITCS6190/Project/DecisionTree/DecisionTree.zip
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/GroupActivity08_Spark_MLlib_AWS.txt

Find instructions to setup the project in
http://webpages.uncc.edu/aatzache/ITCS6190/Project/DecisionTree/README.txt
//write a program , which implements the algorithm from this exercise . use the same data from the exercise as an input, and check your output to match the results of the exercise . user should be asked to provide minimum tree depth treshlod value before program starts .

Group 7

Presentation Subject: Computational Advertising | Comparison between MapReduce and bulk-synchronous systems
Chapter 8. from book 1. MiningOfMassiveDatasets
implement: Graph Analysis in Spark GraphX as described in :
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/SparkGraphX/Exercise_SparkGraphX.txt

Group 8

Presentation Subject: Support Vector Machine
https://webpages.uncc.edu/aatzache/ITCS6190/Project/DM_04_4.9_Chap4_SVM.ppt
implement: Support vectore Machine - Classification using the given data in the project folder.
https://webpages.uncc.edu/aatzache/ITCS6190/Project/SVMInstructions.txt
https://webpages.uncc.edu/aatzache/ITCS6190/Project/SVM.zip

Group 9

Presentation Subject:
Chapter 2 Data from Data Mining Book
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/chapter2_Data.ppt

Implement Exercise 19 chapter 2 in Java and in Spark. Make sure your code produces the correct result as given in exercise 19 solution.
Implementation Links:
1)Java Code
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/Exercise19_Chapter02_SimilarityUsingVectors_JAVA.zip
2)Scala Code
http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/Exercise19_Chapter02_SimilaritiesUsingVectors.zip

The program calculates distance / similarity between animals . Shows how similar is one animal to another.

Group 10

Presentation Subject: Decision Rules (LERS), Action Rule Discovery

Read powerpoints on your research subject rough sets and Action Rules
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/LERS.doc
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/ActionRules_Simple.ppt
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/ActionRuleDiscoveryExample.doc

Implement:
LERS algorithm as described in:
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/08_Exercise_SparkLERS_AWS.txt
and
Action Rules as described in:
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/GroupActivity04_ActionRules_Part4_Spark_AWS.txt

Group 11

Presentation Subject: Clustering
http://webpages.uncc.edu/aatzache/ITCS6162/PowerPoints/KMeansExample.doc

Chapter 7. from Book 1. MiningOfMassiveDatasets

Implement:
K-Means Clustering as described in
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/GroupActivity05_KMeansSparkMLib_AWS.txt

Group 12

Presentation Subject: Classification
http://webpages.uncc.edu/aatzache/ITCS6190/PowerPoints/IR/IR_13_NaiveBayesClassification_Intro.ppt
http://webpages.uncc.edu/aatzache/ITCS6190/PowerPoints/IR/IR_13_TextClassification_NaiveBayes.pptx

Chapter 13. from book 4. Information Retrieval

Implement: Naive Bayes Classification in Spark
Dataset: Car Evaluation and Mammographic dataset
https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/Group_act_12_Naive_Bayes.txt

* Note:
This is a Group Project . On Canvas locate your Group Members , and obtain their e-mails . This project requires that every student checks his/her UNCC e-mail account, and communicates with his / her group-mates . Contact your group-mates as soon as possible . Be sure to talk to them , meet with them , e-mail , telephone , Facebook or use any other means of communication you like . If a student is reported by his / her group-mates as non-responsive or not participating in the group activities , the student will receive a grade of 0 for this project . If a student is not present (misses the class) on the assigned presentation date , the student will receive a grade of 0 for this project .