# AC41011/AC51047: Advanced Big Data Analysis **Big Data Processing Assignment**

**Assignment Description:** Given are three problems. Write up a report giving code for solutions of the problem, as well as an explanation of how your code works. In the explanation, write about the reasoning you have employed in solving the problems, as well as about what each part of your code does. Note that if you **only** submit the code without any explanations, you will be severely penalised in terms of the actual grade that you will obtain.  
  
**How You Will Work on the Problems:** This is an **individual** assignment. You are not allowed to collaborate with anyone else on solving these problems, nor to copy the code from elsewhere on the internet. You are allowed to use the lab sheets for solving the problems. If you use help from any other material (e.g. something from the internet), you need to reference it. If your solution is found to be too similar to someone else’s, or to a piece of code from the internet, this might trigger an academic misconduct investigation.   
  
**How Will Your Submission Be Graded:** Grades will be determined by evaluating the overall quality of the report, as well as on the correctness of Python code that you supplied. Note that solving all the problems and answering all the questions specified will give you the grade A5. To get a better grade, you will need to present an especially elegant/good solution, or you will have to do extra work (of your own choice) in addition to the specification.   
  
**What We are Looking for in the Report:** Good explanation of the reasoning that led you to the solution, as well as of any design decisions that you have made. Where answer to a question involves Python code, the code should be supplied and explained.   
  
Problems

1. **Scala Programming (20%)**. Write a Scala program that reads in a file (it can be either given as an argument to the main method, or hardcoded into it) that counts the number of occurrences of each letter a-z in the file and prints these counts on the screen.  
     
   **Note.** The same capital and lowercase letter should be treated as the same letter. For example, the output for a file that contains just one line  
     
   This is one line of a file  
     
   should be something like  
     
   t:1, h:1, i:3, s:2, o:2, n:2, l:2, e:2, f:2, a:1  
     
   The order of letters in the output does not matter.
2. **Frequent Pattern Mining (40%)**. Given is a dataset in the table Bakery.csv that contains transactions in the bakery. Each row in the transaction contains transaction identifier, item bought, date and time, daytime (morning/afternoon) and daytype (weekend, weekday).   
     
   Write Scala or Python code for Spark that will derive frequent itemsets of bakery items that are frequently bought together and association rules from this data set.   
     
   **Note.** Experiment with different values of minimal support and minimal confidence until you are happy with the resulting recommendations.  
     
   **Note.** You can write a standalone Spark application, or give a series of commands that need to be executed in one of the Spark shells to solve the problem.
3. **Classification** **(40%).** Iris dataset (Iris.csv) is a dataset that contains measurement of iris flowers and their species classification. Write code in Scala or Python for Spark that will use Decision Tree Classification for predicting the flower species based on the four measurement of iris flowers.

**Note**. You can write a standalone Spark application, or give a series of commands that need to be executed in one of the Spark shells to solve the problem.