CST8390 - Lab 2
Data Preparation and Cleaning

Due Date:

- Submission: check Brightspace for submission due date
- Demo: Demo lab during the following week after lab submission

Introduction

The goal of this lab is to clean and prepare data which is in the csv file.

Part 1

1. Download EmployeesSalary.csv file from Brightspace.
2. Open EmployeesSalary.csv in excel and explore it.
3. Read https://www.cs.waikato.ac.nz/ml/weka/arff.html to find the expectations of an ARFF file.
4. Identify the attributes of the data. Record the attributes and the type of attribute for the data.
5. Closely analyse data. In excel, do the required modifications to match with the requirements for an ARFF file. (Hint: Check the requirements if the data has spaces in it.)
6. Open the file in notepad++. Add the required section headers and corresponding information in the file. Save the file as EmployeesSalary.arff. This involves creating the @relation line, one @attribute line per attribute, and @data to signify the start of data. It is good to add comments at the top of the file describing where you obtained this data set, explanation about your attributes etc. A comment in the ARFF format starts with the percent character % and continues until the end of the line.

7. Open the ARFF file as you did in lab 1 (by selecting 'Open file' in the 'Preprocess tab'.). You may run into errors as you load your ARFF file. If so, check the requirements to troubleshoot your problem.

8. Which are the four important attributes that are relevant for data analysis?
   a.

   b.

   c.

   d.

9. For the nominal attributes of the previous question, fill in the following table:

Attribute Name:		**Attribute Name:**	
Label	**Count**	**Label**	**Count**

Attribute Name:	
Label	**Count**

10. Analyze your data to see any anomalies. List the identified anomalies below. Write why you think those records are anomalies in the following format:

Id	first_name	last_name	email	Address	Country	Branch	Currency	Salary	Reason

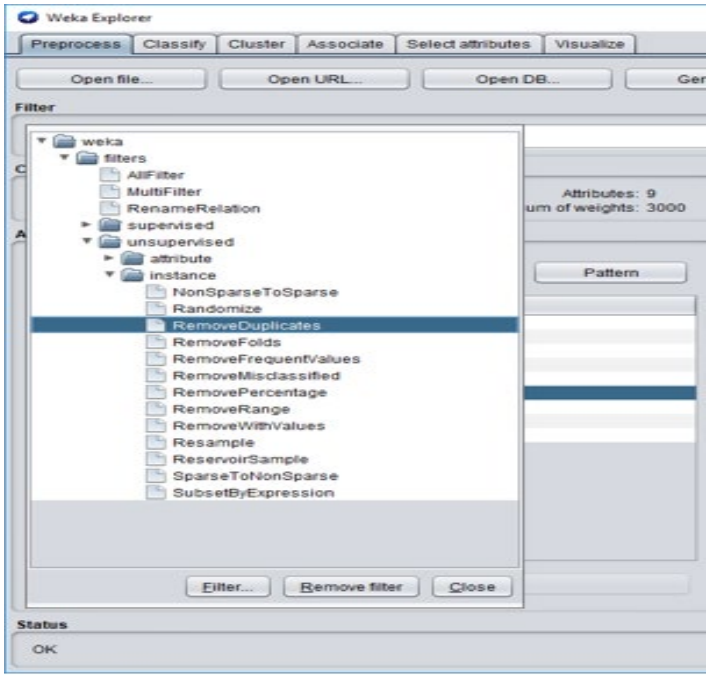
Part 2

Steps:

1. Download EmployeesSalaryBigFile.csv file from Brightspace.
2. Open EmployeesSalaryBigFile.csv in excel and explore it.
3. Identify the attributes of the data. Record the attributes and the type of attribute for the data.
4. Load the CSV file into Weka by selecting 'Open file' in the 'Preprocess tab' (Select CSV data files for the file type).
5. Check different attributes including Branch. Branch is considered as numeric by default. Save the file as arff file by clicking on Save on the right corner.
6. Open EmployeesSalaryBigFile.arff file in notepad++. Change the attribute types of first_name, last_name, email, address, Address and Branch with the required types. Save the file. (This can also be done by applying filters).
7. Open the file again in Weka. Check all attributes and their values.
8. How many instances do you have now? _____
9. Take a screenshot and save it in lab2_Answers.doc.

Remove Duplicates:

10. Check manually whether any duplicates exist in the file.
11. Now run RemoveDuplicates filter to remove duplicates. To do this, from 'Filter', choose weka→ filters → unsupervised → instance → Remove Duplicates.



12. Select Apply to run the filter operation.

13.
    a. How many instances do you have now? _____
    b. How many duplicates (how many got removed): _____
14. Take a screenshot and paste it in lab2 document.
15. Save this new file as EmployeesSalaryBigFileNoDuplicates.arff.

Nominal to Binary

16. How many nominal attributes do you have?
17. With those nominal values, we cannot apply any of the distance-based classification methods. Convert them into binaries using NominalToBinary filter. For that, from Filter, select weka → filters → unsupervised → attribute → NominalToBinary, and hit Apply. (Check other available filters too. You need to use these filters in the future labs).
18. Take a screenshot and paste it in lab2 document.
19. Save this file to EmployeesSalaryBigFileNoDupBinary.arff
20. Open the file in notepad++ and see the data.
21. Take a screenshot of the file while it is opened in Notepad++. Header should be visible.

In order to get the credit for this lab:

1. Show the EmployeesSalary file in Weka during demo.
2. Show EmployeesSalaryBigFileNoDupBinary.arff in Weka
3. Show the answers of Q8,9,10 of part 1 and Q8, 13, 14, 16, 18 & 21 of part 2 in lab2_Answers.doc
4. Upload lab2_Answers.doc in Brightspace before the submission due date.

Both demo during lab hours and submission in Brightspace are required to get credits for the lab.