CST8390 - Lab 3
Data Preparation and Cleaning

Due Date: Week 3 in corresponding lab sessions.

*NOTE: According prof. Anu, it is necessary to show the results **during** LAB sections.*

Introduction

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

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 a word document named Lab3.

Remove Duplicates:

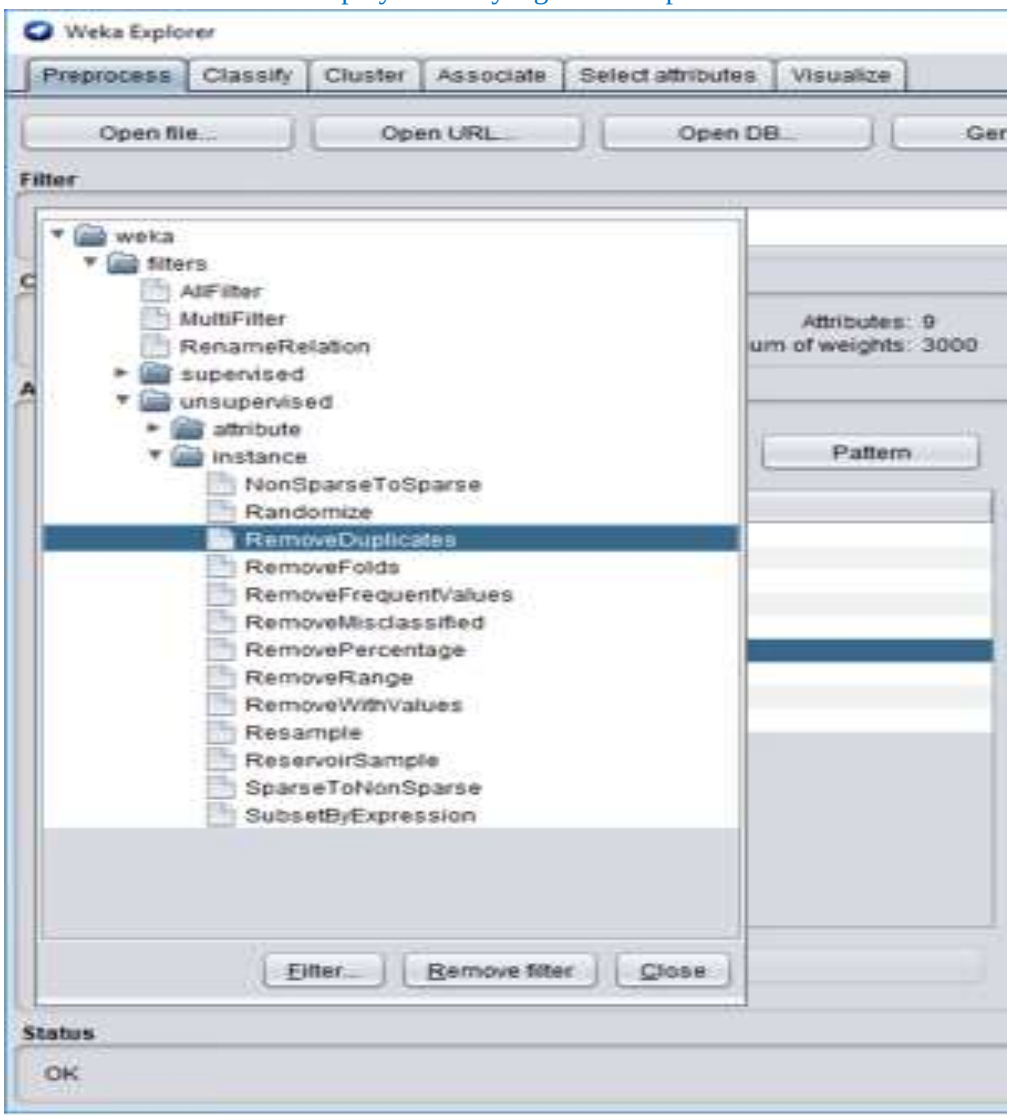
10. Check manually whether any duplicates exist in the file.
11. Now run RemoveDuplicates filter to **remove duplicates**. To do this, from 'Filter', choose weka→ filters → unsupervised → instance → Remove Duplicates.

12. Select Apply to run the filter operation.

13. How many instances do you have now? _____ Duplicates: _____

14. Take a screenshot and paste it in Lab3 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.

18. Take a screenshot and paste it in Lab3 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.

REMEMBER:

In order to get the credit for this lab:

1. Show the screenshots of **Q14**, **Q16** & **Q21** (2 marks);
2. Show EmployeesSalaryBigFileNoDupBinary.arff in **Weka** (3 marks).

FOR YOUR ANALYSIS:

*About the importance of **transforming data** (ex: nominal to binary, string to nominal, etc.) or **removing data**:*

*- In which **circumstances** you should perform these operations and **why**?*

*- Give additional **examples**.*

Ottawa, Jan 2020.