ALGONQUIN COLLEGE

CST8390 - LAB
BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 2

LAB 2 - Explore CSV file and transform it into ARFF file

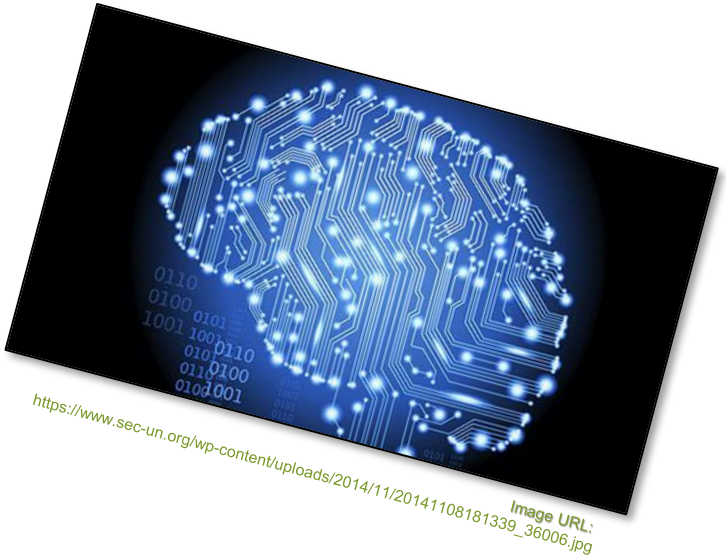
Lab 2 – ARFF files

PART I

- More about ARFF

PART II

- Steps
- Results



CST8390 - Lab
Business intelligence & data analytics

Lab 2 – ARFF Files

Part I – CSV and ARFF

Reviewing CSV

- CSV = Comma Separated Values
  - Text files that can be opened in spreadsheets;
  - The basic idea is to create a “plain” database, using commas to separate attributes (columns) and new lines to identify new entities (rows).

The main idea is that, as any kind of “markup” language, you can create databases easily.

Remember: “Copy and paste” will not work to open in a Excel spreadsheet. You need to save it previously as a CSV file.

Reviewing ARFF

- ARFF = Attribute Related File Format
  - You still need to use CSV format for **@data** section.

Remember: ARFF syntax:

- %: line comment (useful for documentation)
- @relation: name of the “database”
- @attribute: identification of the fields and their datatypes;
- @data: information to be processed (using CSV).

TIP: See https://www.cs.waikato.ac.nz/ml/weka/arff.html

Be careful…

- Sometimes you need to be careful:
  - When copying and pasting information;
  - When defining the attributes and datatypes.

Remember: The <***datatype***> can be any of the four types currently (version 3.2.1) supported by Weka:

1. Numeric (real / integer)
2. Nominal: {<nominal-name1>, <nominal-name2>, ...}
3. String: "*<string>*"
4. Date: ["yyyy-MM-dd'T'HH:mm:ss"]

WARN: Errors in format will stop the loading!

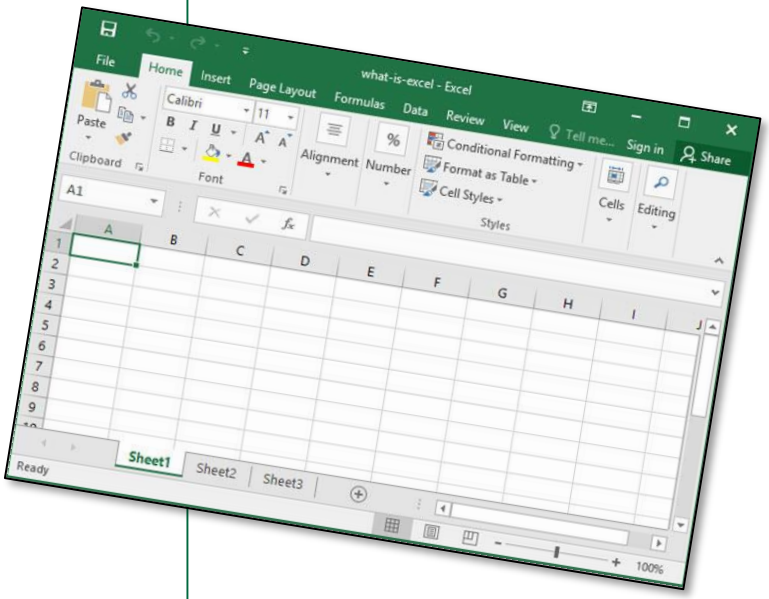
Image URL:https://i0.wp.com/opensourceforu.com/wp-content/uploads/2014/03/Weka-bird.gif?fit=500%2C298&ssl=1

Weka Introduction

Demo

Step-by-step

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.



Step-by-step

4. **Identify** the **attributes** of the data. Record the attributes and the type of attribute for the data.

5. Closely **analyze 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.)

Step-by-step

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 is started with the percent character % and continues until the end of the line.

Step-by-step

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. **QUESTION:** Which are the **four important** attributes that are relevant for data analysis?

Analysis: How did you choose what is and what is not relevant? What is your criteria?

Step-by-step

9. For the nominal attributes from Question 8, **fill** in the following table:

Attribute Name:	
Label	**Count**

Step-by-step

10. Analyze your data to see any anomalies. List the identified anomalies below (you will see at least **8**).

Analysis: How did you consider that you are dealing with an “anomaly”? How to define it?

Get the marks…

In order to get the credit for this lab:

1. Show the loaded file in Weka
2. Fill in the tables for questions 8 & 9
3. Show the list of anomalies

Remember: Include a minimal analysis in the end.

Image URL:
https://storage.googleapis.com/tune-me-bucket/images/A-kiss-for-good-marks.width-280.jpegquality-70.jpg

What about my analysis?

- The previous questions can help you to do your own analysis;
- For instance:
  - What can be considered relevant and WHY?
  - What is an anomaly? How to detect it?
  - Which strategy to use to deal with the correct data?
  - **Use your creativity...**

Tip: The better idea comes from your own analysis. It is your differential.

Image URL: https://www.marketingdirecto.com/wp-content/uploads/2018/01/ciencia-datos.jpg

Open questions…

- Before we start, do you have any doubt / question?

Image URL: https://cdn-az.allevents.in/events3/banners/26b150363d5757da8578fa6a1481585a368f12d4247adfe95837e6ea6c5ab2af-rimg-w1200-h549-gmir.jpg?v=1569691155

See you…

- Remember:
  - Labs require practice and it is ok committing errors and learning with them.
  - Do not forget to show your results…
  - Any questions, let me know…

sousap@algonquincollege.com

Image URL: https://thumbs.gfycat.com/MaleFrigidBull-size_restricted.gif

Thank you for your attention!