**Drexel University**

**College of Computing and Informatics**

**INFO 371 – Data Mining Applications**

**Assignment 3**

**Due Date: Sunday Nov. 3, 2019**

**A. Requirements**

**TEAM (up to three members) assignment**: Please work with up to two other students on this assignment. Please inform the instructor of your team membership before working on the assignment.

# Tools

**-----------------------------------**

* **Weka**: for data pre-processing and data mining
* **Microsoft Excel**: for data pre-processing
* **Text Editor**: for data pre-processing

# Task 1: Mining Association Rules on a Given Data Set (30 points)

**----------------------------------------------------------------------------------------------**

1. Go to your weka installation folder. Under the installation folder, there is **data** folder. Open the **data** folder and find the data set **supermarket.arff.**
2. Open **supermarket.arff** in Weka Explorer.
3. To see the original dataset, click the Edit button, a viewer window opens with dataset loaded.
4. Each row is a transaction. Each column is an item or department number. We are interested in a market basket analysis by finding association rules among the columns/attributes.
5. To do market basket analysis in Weka, each transaction records whether an attribute appears, labeled as ‘t’. Otherwise, the corresponding cell is treated as a missing value.
6. Click **Associate** Tab on top of the window.
7. Choose **weka->association->Apriori**
8. Note that Apriori algorithm expects data that is purely nominal: If present, numeric attributes must be discretized first.
9. Left click the field of Associator. The property window of Apriori opens.
10. Weka runs an Apriori-type algorithm to find association rules.
    1. The min. support is not fixed. This algorithm starts with min. support as upperBoundMinSupport (default 1.0 = 100%), iteratively decrease it by delta (default 0.05 = 5%). Note that upperBoundMinSupport is decreased by delta before the basic Apriori algorithm is run for the first time.
    2. The algorithm stops when lowerBoundMinSupport (default 0.1 = 10%) is reached, or required number of rules – numRules (default value 10) have been generated.
    3. Rules generated are ranked by metricType (default Confidence). Only rules with score higher than minMetric (default 0.9 for Confidence) are considered and delivered as the output.
    4. If you choose to show the all frequent itemsets found, outputItemSets should be set as True.
11. Click Start button on the left of the window, the algorithm begins to run. The output is showing in the right window.
12. You could re-run Apriori algorithm by selecting different parameters, such as lowerBoundMinSupport, minMetric (min. confidence level), and different evaluation metric (confidence vs. lift), and so on.
13. Copy and results to your report. For each result, comment the rules and discuss how their support and confidence values change and are related. Write a paragraph discussing the main findings of your investigation.

# Task 2: Pre-processing Dirty Data for Association Rule Mining (40 point)

**----------------------------------------------**

In this task you will work on a Kaggle data set to prepare it for a market analysis by mining association rules. The data set is available at Kaggle.com: <https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery> In this task, you need to transform the sales data to transactional data for association rule mining.

1. Download the CSV data from the above Kaggle link as **BreadBasket\_DMS.csv**. You may need to sign up an account in Kaggle to download the file.
2. Try to load the original BreadBasket\_DMS.csv to weka. If you see error messages, you need to clean up the data.
3. Open the BreadBasket\_DMS.csv in a plain text editor.
4. Replace all the single quotation marks under the **Item** column with empty strings so that weka can load the CSV data. If it still doesn’t work, continue to clean up the data until weka can load it.
5. Click the Edit button to see the raw data. This data contains 4 columns. Each row records the date, the time, and the item for a transaction (identified by a number). This data format is not good for association rule mining. We need to convert it to a transactional table.
6. Save the CSV file in weka. Open the saved CSV file in Excel.
7. We will use Excel to create a pivot table for the values under the columns **Transaction** and **Item**.
8. On Excel, select the two columns Transaction and Item. Click the **Insert** tab. Click the **PivotTable** icon in the tool bar. You will see a “Create PivotTable” dialog. Make sure the range is selected correctly. Click OK.
9. On **PivotTable Analyze** tab, there is PivotTable Fields area. Check **Transaction** and **Item**. Then drag **Transaction** to the Rows area and drag **Item** to the Columns are. In the Values area, you should see “Sum of Transaction”.
10. Click the Information sign next to “Sum of Transaction”. You will see a PivotTable Fields dialog. Select “Count” instead of “Sum” on the dialog. Click OK.
11. Save the active pivot table sheet at **BreadBasket\_DMS\_pivot.csv**.
12. Open BreadBasket\_DMS\_pivot.csv in Excel again. Remove all the rows before the row with all the attribute names. Change the fisrt column name from “Row Labels” to “Transaction”. Remove the last column “Grand Total” and other blank columns. Remove the second column “Adjustment”. Remove the last row “Grand Total” and other blank rows.
13. Save the cleaned-up CSV file **BreadBasket\_DMS\_pivot.csv.**
14. Load **BreadBasket\_DMS\_pivot.csv** in weka. If you see an error message “null problem encountered at line 9xxx”, open the file in a plain text editor and add a new line at the end of the file.
15. Once the **BreadBasket\_DMS\_pivot.csv** is opened in weka, click the Editor button to see the raw data.
16. On the visualization area, select Transaction (num) as the class.
17. Save the CSV file as ARFF file with the name **BreadBasket\_DMS\_pivot.arff.**
18. Congratulations! You have pre-processed a dirty data set for market analysis. Each row is a single transaction recording what items in the transaction.

# Task 3: Mining Association Rules for the Bread Basket Data (30 point)

**----------------------------------------------**

In this task you will mine association rules on the data set you just created in Task 2.

1. Open **BreadBasket\_DMS\_pivot.arff** in weka and select Transaction as the class.
2. On Filter, choose **weka->filters->unsupervised->attribute->NumericToBinary** -R first-last
3. Click Apply.
4. Check the Transaction attribute and remove it.
5. Click **Associate** Tab on top of the window.
6. Choose **weka->association->Apriori**
7. Click start by using the default settings.
8. Copy the results to your report. Did you find any rules? Why?
9. Re-run the A-priori with changed settings. Test different values on the confidence constraints until you find some rules. Copy the results to your report. Discuss with what values you can find rules and why. Continue to discuss any interesting findings of your investigation.

**B. What to Hand In**

1. Submit your transactional data file **BreadBasket\_DMS\_pivot.arff**
2. A **well-structured** Microsoft Word document or PDF including the following items:
   1. Your name
   2. Course number and title
   3. Assignment number
   4. **The following content in detail**:
      * Step-by-step tasks and requested results.
      * Explain the processes and results so that a reader outside of this class still can understand what you are doing here.
      * Requested comparisons and discussions.
      * Any conclusions must be backed up by the evidence drawn from your data and analyses.

**C. How to Hand In**

1. Please name your report file as **INFO371-assign3-yourFirstName-yourLastName.docx**.
2. Submit your report file through the course website in the **Blackboard Learn** system.

**D. When to Hand In**

1. Submit your assignment no later than **11:59pm** in the due date.
2. There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

**E. Written Presentation Requirements**

Images must be clear and legible. Assignments will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be either 11 or 12 point.