CS 634 Data Mining

Midterm Project

Yasser Abduallah
Department of Computer Science
New Jersey Institute of Technology

Submission Rules

- Embed your last name and first name in your project file name. For example, if your name is John Smith, your file name should read: smith_john_midtermproj.doc or <name>.pdf or <name>.zip, <name>.tar Make sure to include the source code of your project as well as document any additional packages required to run your program.
- Your project will automatically lose **10** points if the above submission rules are violated.
- This is an individual student project.

- Submit your project file in Canvas under Midterm Project Submission Site before the due time. The project file in Canvas is considered as the final version.

- No late project is accepted. A project is late if it is not submitted in Canvas before the due time. Zero points will be given to the late project.

Midterm Project

1. Create 10 items usually seen in Amazon, K-mart, or any other supermarkets (e.g. diapers, clothes, etc.).
2. Create a database of 20 transactions each containing some of these items. The information can be stored in a file, or a DBMS (e.g. NJIT ORACLE or MySQL).

3. Repeat (1) by creating 4 additional, different databases each containing 20 transactions.

Using the Apriori algorithm, generate and print out all the association rules and the input transactions for each of the 5 transactional databases you created. The support and confidence *must* be user-specified parameters, so the output should show different rules with respect to different databases and different support/confidence.
Make sure to show multiple support and confidence results.
The items and transactions must be clear and easy to identify.
Important Notes:

1) The purpose of this project to help you understand the algorithm, therefore, it must be "*your own*" implementation of the algorithm. If you use any existing package for the algorithm from Python, R, Matlab, or Guava, etc…, you will lose points. **For example: Some software has a one-liner function for this algorithm, do not use them for your implementation, but you can use them to verify your work!.**
2) Of course, you can use helper libraries, like pandas, numpy, etc to help you develop the algorithm.
3) *Do not share or copy code from your peers or other resources. Your task is to implement the algorithm from scratch.*

Platforms

- ✓ **Programming language**:
  This is a **Python** based project. If you want to use different programing language you must consult with me first.

- ✓ **Operating system is open:**
  Any one of the following is allowed: Windows, Linux, Mac OS, Ubuntu etc.

- ✓ **Hardware is open:**
  Any one of the following is allowed:
  PC, Laptop, Mac etc.

Project Grading

- The grades will be posted on Canvas when they are completed.

- Note: Keep your project files size small to avoid any problem that may occur when submitting the file in Canvas due to any space limitation.

- The project file must contain the source code and documentation including **screenshots**. The screenshots are used to demonstrate the running situation of your program, particularly how the program executes and produces output based on different input data and user-specified parameter values.

- Implementation, complexity of the code, code style, clarity of the report, and more are taking into consideration.

- Github & Jupyter Book.
    - After you finish your code in development and testing and make sure it works, and prepared the report (meaning all heavy lifting job is done 😊), Create a Github repository in https://github.com/. Your account must be with your NJIT email not your personal email.

- Load your project to the repository.
- Create Jupyter book for your work to show the output, for more info visit https://jupyter.org/
- Give me ya54@njit.edu access as a collaborator to your repository. (If we have a grader, you give him/her access too).
- Add Github link to your repository to your report. NOTE: If you need help with Github and/or Jupyter book, let me know.

NOTE: Jupyter can be used as an IDE to edit and develop your project, but you must save it as Python at the end to generate the source code. Jupyter source code will not be accepted as the project source code.

❖ Project milestone is a mid-way to show your progress. Submit a milestone report that includes:

a. What software version you are using, libraries, etc...
b. Where did you download/get the data (e.g., Amazon site, Kmart site, etc.).
c. Hardware configuration.
d. Did you start coding or not.
e. Any difficulties or issues (bring that to me to discuss).

❖ Copying and sharing code with peers is prohibited and will result in 0 point for all parties that are involved.