

Assessment Description Assessment 2

Assessment name: Association rule/pattern mining for recommender system

Word count/length limit for the report: report approx. 12 pages

Weighting: 25%

CLOs addressed: 1, 3, 4, 5

Purpose: Practice of using association rule mining and recommender system methods.

The goal of assignment 2 is to apply pattern mining and recommendation system methods to solve a practical problem. **This is an individual or a group assignment.** The group may only consist of two or three people. There scope is different for each number of group members.

Using Datasets

Datasets for this assignment contains the following columns:

BillNo	Itemname	Quantity	Date	Price	CustomerID	Cost
bill number of the customer	Item text name	Quantity purchased by this customer this time	Date of transaction	Price of single item	Customer id	Total cost of this transaction

Any or all of these columns can be used for the assignment, your decision which one to use.

Four datasets are provided for this assignment:

- basket_data_by_date_train.csv: to be used for training and tuning.
- basket_data_by_date_test.csv: to be used for obtaining the final results.
- basket_data_by_date_train_big.csv: **can be** used for training a scaled-up system for bigger data if needed. This is not a requirement, just provided for convenience to test scaled-up performance.
- basket_data_by_date_test_big.csv: **can be** used for testing a scaled-up system for bigger data if needed. This is not a requirement, just provided for convenience to test scaled-up performance.

Test transactions are in later dates than training. This is to allow testing “future” recommendations.

Summary of scope and tasks:

1 person: Recommendation system

Download the data, write the code in Python to generate recommendations using collaborative filtering method from the training dataset. Your method should be able to generate recommendations from big data and for large number of users and be scalable. Your method should score recommendations for a user to select top recommendations. Select a metric to measure the performance. Write code to test the recommendations on the test set.

2 persons: Recommendation system from frequent patterns

Person 1: Download the data, write the code in Python to mine frequent patterns from the training dataset. You can use any suitable pattern mining algorithm including those discussed in the course, e.g. Apriori or FP-Growth. Research and use a method suitable for frequent pattern mining for big data. Describe the method and cite references and why the method is selected.

Input: **data**, output: **patterns**.

Person 2: Download the data, write the code in Python to generate recommendations from frequent patterns delivered by Person 1. Research a method of creating recommendations from frequent patterns. Describe the method and cite references. Your method should be able to generate recommendations from a set of patterns. Your method should score recommendations for a user to select top recommendations. Select a metric to measure the performance. Write code to test the recommendations on the test set.

Input: **patterns**, output: **recommendations**.

3 persons: Improved recommendation system from frequent patterns

Person 1: Download the data, write the code in Python to mine frequent patterns from the training dataset. You can use any suitable pattern mining algorithm including those discussed in the course, e.g. Apriori or FP-Growth. Research and use a method suitable for frequent pattern mining for big data. Describe the method and cite references and why the method is selected.

Input: **data**, output: **patterns**.

Person 2: Download the data, write the code in Python to generate recommendations from frequent patterns delivered by Person 1. Research a method of creating recommendations from frequent patterns. Describe the method and cite references. Your method should score recommendations for a user to select top recommendations. Select a metric to measure the performance. Write code to test the recommendations on the test set.

Input: **patterns**, output: **recommendations**.

Person 3: (suitable for someone interested in Natural Language Processing). Research a method to address cold start problem. One way is to assess item similarity may be based on their item name. You are free to research and use other methods. This will address the cold start problem for items that are in test set but not in training set. Use this method to address the cold start problem in collaborative filtering, and to develop a content-based recommender from item similarity. Combine methods from person 1 and 2 (CF and contents based) with the goal to improve recommendations. Show if and how your method improves recommendations compared to recommendations without your method.

Report structure.

The main body of report should be limited to 12 A4 pages including references, but excluding any appendices, noting that including **necessary** contents has a priority over page limit. The main body of the report should contain minimum information for the reader to understand your project in detail. Any extra information, charts and tables, that take too much space, should be placed in the appendix, and referred from the main body of the report.

The report should contain:

1. **Title page:** title of the project, names and ids of group members.
2. **Executive summary** (non-technical, ≤ 1 page). This section is for the company management that may not be familiar with technical details. It should include a brief problem description, benefits for the company and feasibility of scaling the solution. You can include test results, but need to explain what they mean for a layperson.
3. **Introduction (this starts the technical report):** a brief explanation of the problem, the aim of the project, and expected business benefits.

4. **Exploratory analysis:** analysis of data that will give some insights how to use it, and potential solutions and potential problems that you may encounter. Include charts and tables of the analysis as needed. Each part of analysis, charts and tables, should have its purpose why you decided to include it and how it is useful.
5. **Implementation part should include:**
 - a. Diagram and description of the whole system
 - b. Diagram and description of each part of the project (if the team is more than one person)
 - i. If you use an algorithm or method not discussed in lecture or workshop, describe the method and include a reference that gives more information.
 - c. Training/testing/evaluation methodology.
 - i. Train and tune on the training set, use test set for the final results.
 - ii. How the results were obtained, what metrics were used for evaluation. How the patterns and/or recommendations were ranked.

6.

Discussion of results (Options 1).

- i. Five examples of recommendations that the user actually bought as evidenced in the test set.
- ii. Table or chart of metrics with discussion, showing results of testing recommendations on the test set.

Discussion of results (Options 2 and 3).

- b. Five examples of frequent patterns with their support in both training and test sets.
- c. 10 examples of recommendations: two examples from each of the above patterns.
- d. Table or chart of metrics with discussion, showing results of testing frequent patterns on the test set.
 - i. If the system has pattern mining and recommendation system (2-person team), show and discuss result of recommender performance with and without pattern mining.
 - ii. If the system has pattern mining and recommendation system and improved recommendation system (**Option 3: 3-person team**), show and discuss result of recommender performance with and without NLP improvement (that is, no need to show performance with and without pattern mining)..
7. **Conclusion and Recommendations:** which method do you recommend for the end user and why. Include scaling up consideration and benefits for the company. Include future improvements.
8. **Reflection:** what is one main thing (if any) you have learned through this project and what would you do better next time.
9. **References (Harvard)**
 - a. If no references are used in the assignment (very unlikely case), please make the following statement: **"No references used."**

Practical tips:

1. If the training/testing time takes too long (e.g. more than a couple of hours) using your available resources, then you can use a subset of the datasets, preserving the proportion of training/test. In this case, make a comment in the implementation part of the report.
2. Always start training/testing on a much smaller dataset that allows quick insights and debugging.

Submission details

Group report and final code.

- Do not include the dataset in the submission.
- Please submit two files (do not zip):
 - One code file “<list_of_ids>_assign2.ipynb”
 - Report as pdf file “list_of_ids>_assign2.pdf”.
(for example a1234567_a234567_assign2.pdf)
- Submit **only one final report and one final code for the group.** (added 4/05/2023)

Late submission rules:

- 1 day late – mark capped at 75%
- 2 days late – mark capped at 50%
- 3 days late – mark capped at 25%
- more than 3 days late – no marks available

Submitting the final code is essential to mark the report.

E. Suggested reference papers:

Lee, C.H., Kim, Y.H. and Rhee, P.K., 2001. Web personalization expert with combining collaborative filtering and association rule mining technique. Expert Systems with Applications, 21(3), pp.131-137.
https://www.sciencedirect.com/science/article/pii/S0957417401000343/pdf?casa_token=68QbbDQ8URAAAAAA:cmETEPc3pOkzxZC_vF_dKDgU41Rwlyj1ZpQgGwkvNjIMVyCCZLGNYWUVvPr-HZCNwSDrAhGGIA&md5=c9cc1e9206467a1b3df9d70ad12f93b7&pid=1-s2.0-S0957417401000343-main.pdf

Parvatikar, S. and Joshi, B., 2015, December. Online book recommendation system by using collaborative filtering and association mining. In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICIC) (pp. 1-4). IEEE.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7435717>

Academic Integrity Declaration

By submitting this assignment, I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I acknowledge that the assessor of this item may, for the purpose of assessing this item, reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking). I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.