Homework assignment 5:
Association mining

Objective: The overall objective is to understand how frequent itemsets can be extracted by the Apriori algorithm and be able to calculate and interpret association rules in terms of support and confidence.

Material: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, section 6.

Additional tools: For this homework, you can install/load script `runApriori.py` provided on the Blackboard.

Important: The following points are how you hand-in the homework assignment.

- Provide clear and complete answers to the questions below (not hidden somewhere in your source code), and make sure to explain your answers / motivate your choices. Please make as PDF file.
- Source code, output graphs, derivations, etc., should be included, and zipped together with the PDF file.
- Hand-in: upload to Blackboard.
- Include name, student number, assignment (especially in filenames)
- For problems or questions: use the BB discussion board or email.

5.1 Association mining for course data

We will use the Apriori algorithm to automatically mine for associations.The Apriori algorithm we use is provided by `http://www.borgelt.net/apriori.html`. We have generated the script `runApriori.py` that enable to call the algorithm from Python. Look at the script `runApriori.py` to see how the external algorithm can be called from Python environment

5.1.1 (0 points) Load the data file `Data/courses.txt` into Python. The data is represented in Table 1. Inspect the file Data/courses.txt and make sure you understand how the data in Table 1 is stored in the text file.

Table 1: Students that upon completing their engineering degree had taken various of the courses

	History	Math	Biology	Spanish	Economics	Physics	Chemistry	English
student 1	0	1	0	0	1	1	1	1
student 2	1	1	1	0	0	1	1	1
student 3	0	1	0	1	0	1	0	1
student 4	0	0	1	0	0	1	1	0
student 5	0	1	0	0	0	1	1	0
student 6	0	1	1	0	0	1	1	1

5.1.2 (1 point) We will analyze the data in Table 1 automatically using the script `runApriori.py`. Analyze the data with $minsupport \geq 80\%$ and $minconfidence \geq 100\%$.What are the generated association rules? What kind of conclusions can you make based on these association rules about the subjects that students took?

(Notice if you are running Python on a linux computer you need to give permission to execute the file containing the apriori algorithm by typing chmod +x apriori).

5.2 Association mining for MovieLens data

In this part of the exercise we consider a Market Basket data set containing 943 users purchases of 1682 movies. A total of 100,000 movies have been purchased.The data set is called MovieLens100K and is provided by http://www.grouplens.org/node/73, see also the readme `MovieLensData.txt` in the data folder.The data currently considered is not the original data but modified for the apriori algorithm.

5.2.1 (0 points) The MovieLens data is stored in the file Data/MovieLensData.txt. Inspect the file to see how the data is stored.

5.2.2 (1 point) Find association rules using the apriori.m with $minsupport \geq 30\%$ and $minconfidence \geq 80\%$. What are the associations with strongest confidence? Do these associations make sense? You can use file Data/u.item to check the names of the movies.

5.2.3 (1 point) Which movie has been bought by the most users? There are only few rules with more than three items. Why?

5.2.4 (0.5 points) Often we are interested in rules with high confidence. Is it possible for itemsets to have very low support but still have a very high confidence?

5.3 **Paper and pencil exercise**

Do the following paper and pencil exercises by hand (rather than computer) and write down how you computed things, not just the answers themselves:

5.3.1 (3.5 points) Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 45%, the support for item b is 80% and the support for itemset $\{a, b\}$ is 30%. Let the support and confidence thresholds be 20% and 60%, respectively.

i. Compute the confidence of the association rule $\{a\} \rightarrow \{b\}$. Is the rule interesting according to the confidence measure?

ii. Compute the interest measure for the association pattern $\{a, b\}$. Describe the nature of the relationship between item a and item b in terms of the itemset measure.

iii. What conclusion can you draw from the results of parts (i) and (ii)

iv. Prove that if the confidence of the rule $\{a\} \rightarrow \{b\}$ is less than the support of $\{b\}$ then

$$c(\{\bar{a}\} \rightarrow \{b\}) > c(\{a\} \rightarrow \{b\}) \tag{1}$$

$$c(\{\bar{a}\} \rightarrow \{b\}) > s(\{b\}) \tag{2}$$

where $c(\cdot)$ denotes the rule confidence and $s(\cdot)$ denote the support of an itemset.

Hint: To prove the statement rewrite the confidence and the support of the rule in terms of probabilities:

$$c(\{a\} \rightarrow \{b\}) = \frac{P(\{a, b\})}{P(\{a\})} \tag{3}$$

$$s(\{b\}) = P(\{b\}) \tag{4}$$

5.3.2 (3 points) Consider the relationships between customers who buy high-definition televisions and exercise machines as shown in Table 2 and 3.

i. Compute the odd ratios for both tables.

ii. Compute the ϕ-coefficient for both tables.

iii. Compute the interest factor for both tables.

For Tables 3 you should compute measures given above separately for College Students and for Adults. For each of the measures, describe how the direction of association changes when data is pooled together (Table 2) instead of being separated into two groups (Table 3).

Table 2: Two way contingency table between the sale of high-definition television and exercise machine

Buy HDTV	Buy Exercise machine		
	Yes	No	
Yes	105	87	192
No	40	62	102
	145	149	294

Table 3: Example of three-way contingency table

Customer group	Buy HDTV	Buy Exercise machine		Total
		Yes	No	
College Students	Yes	2	9	11
	No	5	20	25
Working Adult	Yes	103	78	181
	No	35	42	77