Assignment 3

1. (5 pts) Complete the accompanying python notebook on creating a neural net using PyTorch.

2. (5 pts) Consider the following set of users and movies they have rated.

'Lisa Rose': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.5, 'Just my Luck': 3.0, 'Superman Returns': 3.5, 'You, Me and Dupree': 2.5, 'The Night Listener': 3.0}	'Gene Seymour':{'Lady in the Water': 3.0, 'Snakes on a Plane': 3.5, 'Just my Luck': 1.5, 'Superman Returns': 5.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 3.5}
'Michael Phillips':{'Lady in the Water': 2.5, 'Snakes on a Plane': 3.0, 'Superman Returns': 3.5, 'The Night Listener': 4.0}	'Claudia Puig': {'Snakes on a Plane': 3.5, 'Just my Luck': 3.0, 'The Night Listener': 4.5, 'Superman Returns': 4.0, 'You, Me and Dupree': 2.5}
'Mick LaSalle':{'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0, 'Just my Luck': 2.0, 'Superman Returns': 3.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 2.0}	'Jack Matthews':{'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0, 'Superman Returns': 5.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 3.5}
'Toby': {'Snakes on a Plane': 4.5, 'Superman Returns': 4.0, 'You, Me and Dupree': 1.0}	

(a) **(3 pts)** Suppose we build a recommender system following the user-user similarities approach with Pearson correlation as a similarity measure. What will be the rating prediction for user Michael Phillips, for movie “You, Me and Dupree”? Give the details of your computation.

In computing the Pearson user-user similarities, restrict the user vectors to only those components (movies) the two users have in common.

Consider only the positive similarities for producing the rating prediction.

(b) **(2 pts)** If we use the user-bias, item-bias approach to recommendation (Netflix competition), what will b_r (short for $b_{\text{lisa rose}}$) be after the first pass over the data? Set λ_1=λ_2=γ=0.1, and start with zero bias values.

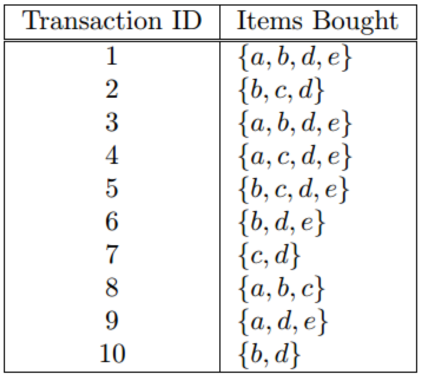
For this question you can assume bi's are all 0, and only update bu for the given user, ignoring bi's at every step.
E.g.
br = 0 %and all bi’s are 0 initially. The predicted r_ui is mu+bu+bi or just mu+bu since bi’s are 0 initially.

br = br + gamma*((2.5-(mu+br))- lambda*br)
-0.072857
br = br + gamma*((3.5-(mu+br))- lambda*br)
-0.037700
...

3. (5 points) Use the following **similarity matrix** to perform MIN and MAX hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. Here you should be careful to not confuse similarity with distance. If you would like to work with distance (similar to the slides), then replace each similarity value x in the table with 1-x.

	p1	**p2**	**p3**	**p4**	**p5**
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

4. (5 points) The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k + 1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in the table with minsup=30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.



Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(a) Draw an itemset lattice representing the data set. Label each node in the lattice with the following letter(s):

N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

F: If the candidate itemset is found to be frequent by the Apriori algorithm.

I: If the candidate itemset is found to be infrequent after support counting.

(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

(c) What is the pruning ratio of the Apriori algorithm on this data set?
(Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

(d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

5. (5 points) Using the data in the above table, build an FP-Tree, and then mine the frequent itemsets using FP-Growth with minsup=30%.