Tuesday Quiz

1. [1 point] Circle ALL of the statements that are TRUE about TF-IDF.
   A. TF-IDF = document frequency * inverse term frequency
   B. Words with highest TF-IDF score are often the terms that best characterize the topic of a document
   C. Most Frequent term has TF = 1
   D. Term that occurs in every document has IDF = 0
   Answer. B,C,D

2. [1 point] Pearson Correlation measures the extent to which two linear variables relate.
   True

3. [0.5+0.5 points]How are ratings normalized? What is the advantage of using normalized ratings?
   Answer. Ratings are normalized by subtracting the mean rating from each rating for a user.
   Advantage of using normalized ratings is that it removes the user bias when it comes to ratings. Some users are hard raters, some are easy raters. Hence normalization gives a common grounds on which ratings from different user-ratings can be compared fairly.

4. [2 points] Give two disadvantages/cons of Content Based Approach with explanation.
   1. Finding appropriate features can be difficult
   2. Recommending items to new users is difficult since there is no user profile
   3. Overspecialization: Items are suggested only from what user may like and no out of the box suggestions are made

5. [0.5+1 Points] Briefly describe the Cold Start Problem in recommender systems? What are the cases that could lead to the Cold Start Problem?
   Answer -When a new user or item has just entered the system, it is hard to find similarities as there is not enough information to make good recommendations.
   New item problem: can't be recommended until some users rate it. Also called "first-rater Problem"
   New users: not given good recommendations because of lack of rating or purchase history

6. [3.5 Points] Consider the following table where rows are the Users and columns are the Items. The values corresponding to ratings.

	I1	I2	I3	I4
U1	2	1		3
U2	3	?	5	2
U3		4	2	3
U4	5	3	1	

Find the recommendation rating on I2 for U2(shown by **?**) using User-Based CF. You need to use Pearson Correlation for your computations.

W(1,2) = -1
Numerator   = (2-2.5) (3-2.5) + (3-2.5)(2-2.5)
Denominator = ((2-2.5)^2+(3-2.5)^2)^0.5 * ((3-2.5)^2 + (2-2.5)^2)^0.5

W(2,3) = -1
Numerator   = (5-3.5) (2-2.5) + (2-3.5)(3-2.5)
Denominator = ((5-3.5)^2+(2-3.5)^2)^0.5 * ((2-2.5)^2 + (3-2.5)^2)^0.5

W(2,4) = -1
Numerator   = (3-4) (5-3) + (5-4)(1-3)
Denominator = ((4-3)^2+(2-3)^2)^0.5 * ((3-2)^2 + (1-2)^2)^0.5

P(2,2) = 3.33 + ((1-2.5)*-1 + (4-2.5)*-1 + (3-3)*-1) / (1 + 1 + 1)
        = 3.33 + 0
        = 3.33

All weights correct - 2 points
Final answer correct - 1.5 points
Marking has been lenient for this question as a lot of you ran out of time.