Take-home quiz

1. [1+2 points] Given a set of documents, briefly explain how to calculate TF and IDF in TF-IDF score. You need to describe any preprocessing you need to apply to the words in a document (open-ended question) and how to calculate both the TF and IDF components.

   Preprocessing steps:
   1. eliminate the stop words, that are common in the docs but not important
   2. Remove punctuations.
   3. Convert to lowercase for uniformity

   TFij = fij / maxkfkj, where
   fij = frequency of term i in document j
   max fkij is the most occurrences of any term in document j

   IDFi = log$_2$(N/ni)
   ni = number of documents containing term i
   N = total number of documents

2. [2 points] Briefly explain one advantage and one disadvantage of using Decision Trees for finding recommendations compared to using the Cosine Distance.

   Advantage: more accurate and works on small problem sign
   Disadvantage: consider difficult prediction/complex combination

3. [0.5+1 points] What is the long tail effect and how do you address this problem in recommendation systems? Give an example other than the ones discussed in class.

   The long tail refers to a phenomenon in which a small proportion of items are really popular (these form the head of the tail) and others, although large in number are unpopular(the heavy tail).
   Downloading music from iTunes; Netflix, etc

4. [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 I4 for U4(shown by **?**) using User-Based CF. You need to use Pearson Correlation for your computations.

W(1,4) = 1
Numerator = (2-1.5) (5-4) + (1-1.5)(3-4)
Denominator = ((2-1.5)^2+(1-1.5)^2)^0.5 * ((5-4)^2 + (3-4)^2)^0.5

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

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

P(4,4) = 3 + ((3-1.5)*1 + (2-4)*-1 + (3-3)*1) / (1 + 1 + 1)
= 3 + 7/6
= 4.167

All weights correct - 2 points
Final answer correct - 1.5 points