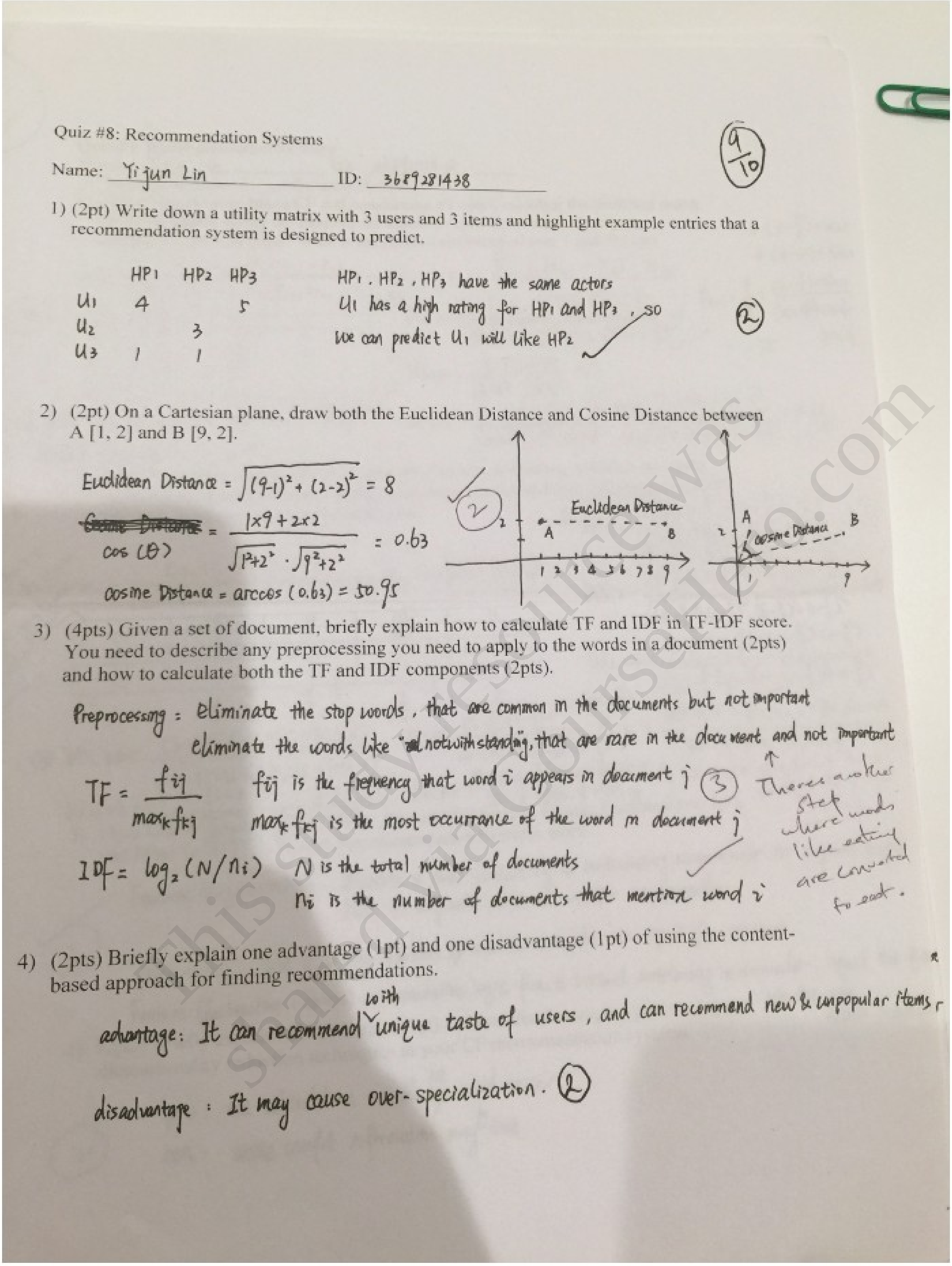

Quiz #8: Recommendation Systems

$\frac{9}{10}$

Name: Yijun Lin ID: 3689281438

1) (2pt) Write down a utility matrix with 3 users and 3 items and highlight example entries that a recommendation system is designed to predict.

	HP1	HP2	HP3
U1	4		5
U2		3	
U3	1	1	

HP1, HP2, HP3 have the same actors
U1 has a high rating for HP1 and HP3, so
we can predict U1 will like HP2 ✓

2

2) (2pt) On a Cartesian plane, draw both the Euclidean Distance and Cosine Distance between A [1, 2] and B [9, 2].

Euclidean Distance = $\sqrt{(9-1)^2 + (2-2)^2} = 8$

~~Cosine Distance~~ cos (θ) = $\frac{1\times 9 + 2\times 2}{\sqrt{1^2+2^2}\cdot\sqrt{9^2+2^2}} = 0.63$

cosine Distance = arccos (0.63) = 50.95

✓ 2

3) (4pts) Given a set of document, 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 (2pts) and how to calculate both the TF and IDF components (2pts).

Preprocessing: eliminate the stop words, that are common in the documents but not important
eliminate the words like "notwithstanding", that are rare in the document and not important

$TF = \frac{f_{ij}}{\max_k f_{kj}}$ f_{ij} is the frequency that word i appears in document j
$\max_k f_{kj}$ is the most occurrance of the word in document j ✓

$IDF = \log_2(N/n_i)$ N is the total number of documents
n_i is the number of documents that mention word i

3

↑ There's another step where words like eating are converted to eat.

4) (2pts) Briefly explain one advantage (1pt) and one disadvantage (1pt) of using the content-based approach for finding recommendations.

advantage: It can recommend with unique taste of users, and can recommend new & unpopular items,

disadvantage: It may cause over-specialization. 2