Text Mining - PRACTICE PROBLEM

Assume you have the following term frequency matrix:

	angeles	los	new	post	times	york
D1	0	0	1	0	1	1
D2	0	0	3	1	0	3
D3	2	2	0	0	1	0

1. Transform the weights using TF-IDF WEIGHTING. Use the traditional Log in the calculations (Log in base 10)
2. Using the transformed matrix, compute the cosine similarity between D1-D2

1. TF * IDF = TF * log(N/ n_w)

Where TF is the Term Frequency

Log(N/ n_w), N is the total number of documents in the dataset; n_w is in how many documents the word under consideration appears

The matrix above already contains TF, since it is a term frequency matrix. SO we already have the first part of TF-IDF.

We need to compute IDF for each word as log(N/ n_w). Let us use the traditional log, that is the log in base 10 for the calculations.

The matrix below shows the calculations for each word. Note that the IDF for each word is going to be the same across documents.

	angeles	los	new	post	times	york
D1	Log(3/1) = 0.477	Log(3/1) = 0.477	Log(3/2) = 0.176	Log(3/1) = 0.477	Log(3/2) = 0.176	Log(3/2) = 0.176
D2	0.477	0.477	0.176	0.477	0.176	0.176
D3	0.477	0.477	0.176	0.477	0.176	0.176

Then multiply TF * IDF: this is going to be different for each word-document cell

	angeles	los	new	post	times	york
D1	0* 0.477= 0.000	0* 0.477= 0.000	1* 0.176 = 0.176	0* 0.477= 0.000	1* 0.176 = 0.176	1* 0.176 = 0.176
D2	0*0.477 = 0.000	0* 0.477= 0.000	3* 0.176 = 0.528	1* 0.477 = 0.477	0* 0.176 = 0.000	3* 0.176 = 0.528
D3	2* 0.477 = 0.954	2* 0.477 = 0.954	0* 0.176 = 0.000	0* 0.477= 0.000	1* 0.176 = 0.176	0* 0.176 = 0.000

2. Calculate the Cosine Similarity between D1 – D2 using the TF-IDF matrix

$$Sim(d_1, d_2) = \frac{\sum_{i=1}^{N} w_{1i} * w_{2i}}{\sqrt{\sum_{i=1}^{N} w_{1i}^2} * \sqrt{\sum_{i=1}^{N} w_{2i}^2}}$$

w1i = weight for the word under consideration in document 1
w2i = weight for the word under consideration in document 2

	angeles	los	new	post	times	york	**Sum**	**Sqrt**
w1i * w2i	0 * 0 = 0.000	0*0 = 0.000	0.176* 0.528 = 0.093	0*0.477 = 0.000	0.176* 0 = 0.000	0.176*0.528= 0.093	0.186	
w1i^2	0.000	0.000	0.176^2 = 0.031	0.000	0.176^2 = 0.031	0.176^2 = 0.031	0.093	0.305
w2i^2	0.000	0.000	0.528^2 = 0.279	0.477^2 = 0.228	0.000	0.528^2 = 0.279	0.786	0.886
Final								**0.688**

Remember that Cosine Similarity is a Similarity Index: as such, the greater the number, the more similar the two documents