Lesson 25 (tf-idf) The idf of a term (word) is equal to:

$$\text{idf(term)} = \ln\left(\frac{\text{total number of documents} + 1}{\text{number of documents containing the term} + 1}\right) + 1$$

Consider the collection of sentences given below.

Document	
0	The car crashed long ago.
1	The car has rusted.
2	A rusted car is unsafe.
3	Spare car parts are needed urgently.

*TFIDF score for term i in document j = TF(i,j) * IDF(i)*
where
IDF = Inverse Document Frequency
TF = Term Frequency

$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$

$IDF(t) = \log\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$

and
t = Term
j = Document

IDF看上面

(a) Compute the idf (inverse document frequence) for each term in the vocabulary. (Remove stop words first.)

Inverse Document Fequency (idf):

terms (words)	ago	car	crashed	long	needed	parts	rusted	spare	unsafe	urgently
# documents	1	4	1	1	1	1	2	1	1	1
idf	1.916	1	1.916	1.916	1.916	1.916	1.511	1.916	1.916	1.916

(b) Compute the tf-idf matrix for the collection of sentences.

tf-idf

tf: term frequency in that document -> remove stop words

normalization: divided by the sum of whole column -> answer will be same as what we got on computer

terms (words)	ago	car	crashed	long	needed	parts	rusted	spare	unsafe	urgently
Document 0	0.48	0.25	0.48	0.48						
1		0.5					0.76			
2		0.33					0.5		0.5	
3		0.2			0.38	0.38		0.38		0.38

(c) Describe what the tf-idf values equal in a column corresponding to a word that occurs in all of the sentences.

(d) What will the tf-idf values equal in a column corresponding to a word that occurs in none of the sentences?