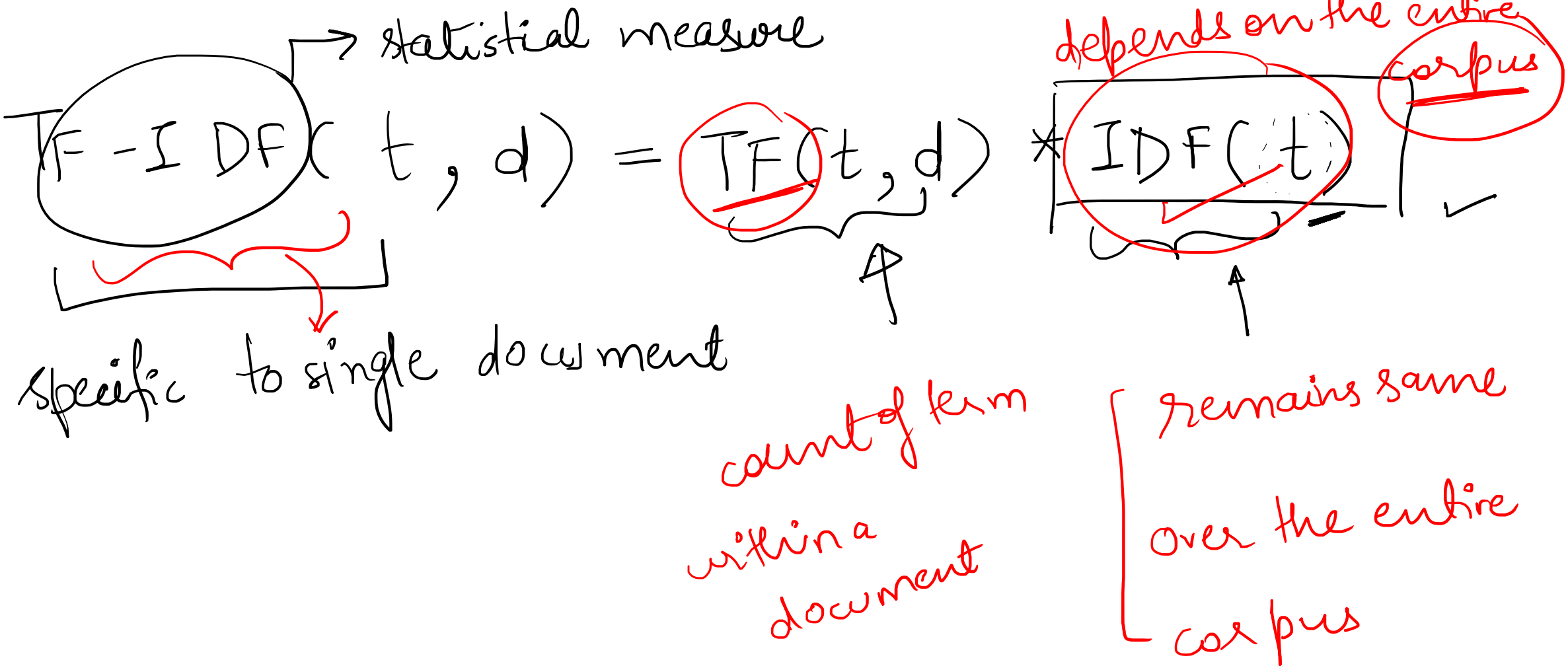
Week 7 NLP Live Session

TF-IDF

TF: Term frequency

IDF: Inverse Document frequency


statistical measure
TF-IDF(t, d) = TF(t, d) * IDF(t)
depends on the entire corpus
specific to single document
count of term within a document
remains same over the entire corpus

↑ term count within the doc.

$$TF(t,d) = \frac{\text{no. of times term 't' appears in doc 'D'}}{\text{No. of terms in doc 'D'}}$$

$$IDF(t) = \log\left(\frac{\text{No. of documents in the corpus}}{\text{No. of doc in which term 't' appears}}\right)$$

Q1: Consider a document 'd' containing 100 words wherein the word 'cat' appears 3 times.

Now, suppose in the corpus there are 10 million (10^6) documents & the word 'cat' appears in 1000 of these doc. Calculate TF-IDF (cat, d)

1 million = 10^6

$$tf(cat, 'd') = \frac{\text{no. of times 't' appears in d}}{\text{no. of terms in d}} = \frac{3}{100}$$

$$= 0.03$$

$$IDF(t) = \log\left(\frac{\text{no. of doc in the corpus}}{\text{no. of doc in which term 't' appears}}\right)$$

$$= \log\left(\frac{10 * 10^6}{1000}\right)$$

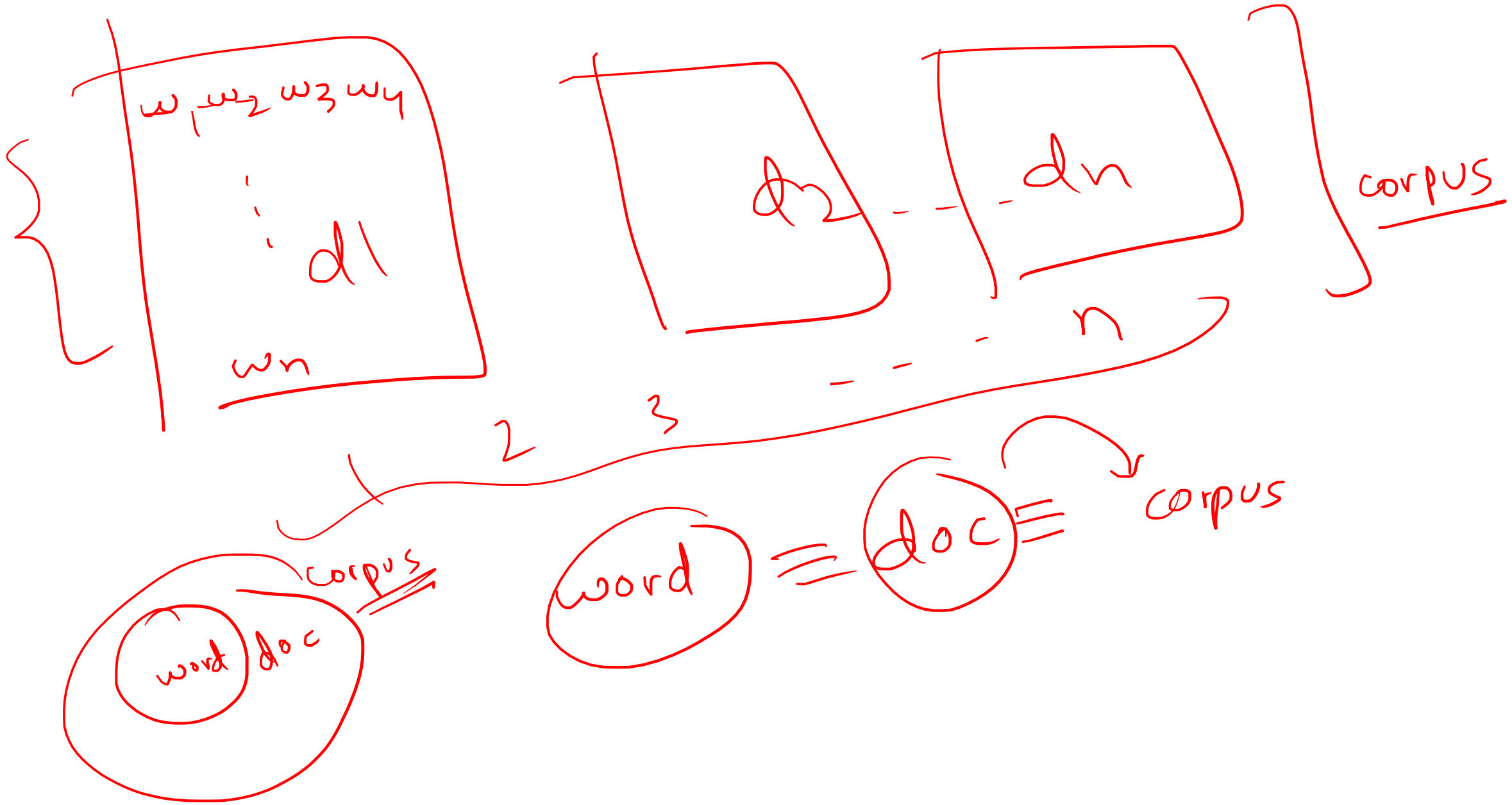
$$= \log\left(\frac{10^7}{10^3}\right) = \log(10^4) = 4 \log_{10} 10$$

$$= 4$$

$$\log_{10} 10^4 = 4$$

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t)$$

$$= \frac{3}{100} * 4 = 0.12$$


w1 w2 w3 w4
d1
wn
d2
dn
corpus
n
2
3
corpus
word
doc
word
doc
corpus

Q2) D1: A quick brown fox jumps over the lazy dog.
what a fox!

|D1| = 12

D2: A quick brown fox jumps over the lazy fox.
What a fox!

|D2| = 12

~~D3~~: How word fox is relevant to corpus 'D' doc?

Hint: TF - IDF (fox, D)

$$TF(fox, D1) = \frac{\text{No. of times fox appears in D1}}{\text{No. of terms in D1 } (|D1|)}$$

$$= 2/12 = 1/6 = 0.17$$

$$TF(fox, D2) = \frac{3}{|D2|} = \frac{3}{12} = \frac{1}{4} = 0.25$$

$D = \{D1, D2\}$

$$IDF(fox, D) = \log\left(\frac{\text{No. of doc in corpus } D}{\text{No. of doc in which 'fox' appears}}\right)$$

$$= \log\left(\frac{2}{2}\right)$$

$$= 0$$

$$TF\text{-}IDF(fox, D1) = TF(fox, D1) * IDF(fox)$$

$$= 0.17 * 0 = 0$$

$$TF\text{-}IDF(fox, D2) = TF(fox, D2) * IDF(fox)$$

$$= 0.25 * 0 = 0$$

Thus: The word 'fox' is equally relevant to corpus D

$\therefore$ TF-IDF(fox, D1) = (TF-IDF(fox, D2) is same

tf: simple choice (raw count of a term in a doc)

idf: how much info^{n} the word provides in our corpus

TF-IDF: statistical measurement

↳ order / sequence of words (~~semantic~~)

Q3: d1: the man went out for a walk.

d2: the children sat around the fire

TF(the, D1) = 1/7, TF(the, D2) = 2/6 = 1/3

IDF(the, D) = log(2/2) = 0

(a) ~~TF~~ IDF (the, D) = 0

(b) TF IDF (fire, D) 1/6 (log 2)

(c) TF IDF (children, D) 1/6 (log 2)

$TF(fire, D1) = 0$ ✓

$TF(fire, D2) = \frac{1}{6}$

$IDF(fire, D) = \log\left(\frac{2}{1}\right) = \log_{10} 2 = 0.3010$

$TFIDF(fire, D1) = 0$

$TFIDF(fire, D2) = \frac{1}{6} * \log 2$

R1: This movie is very scary and long

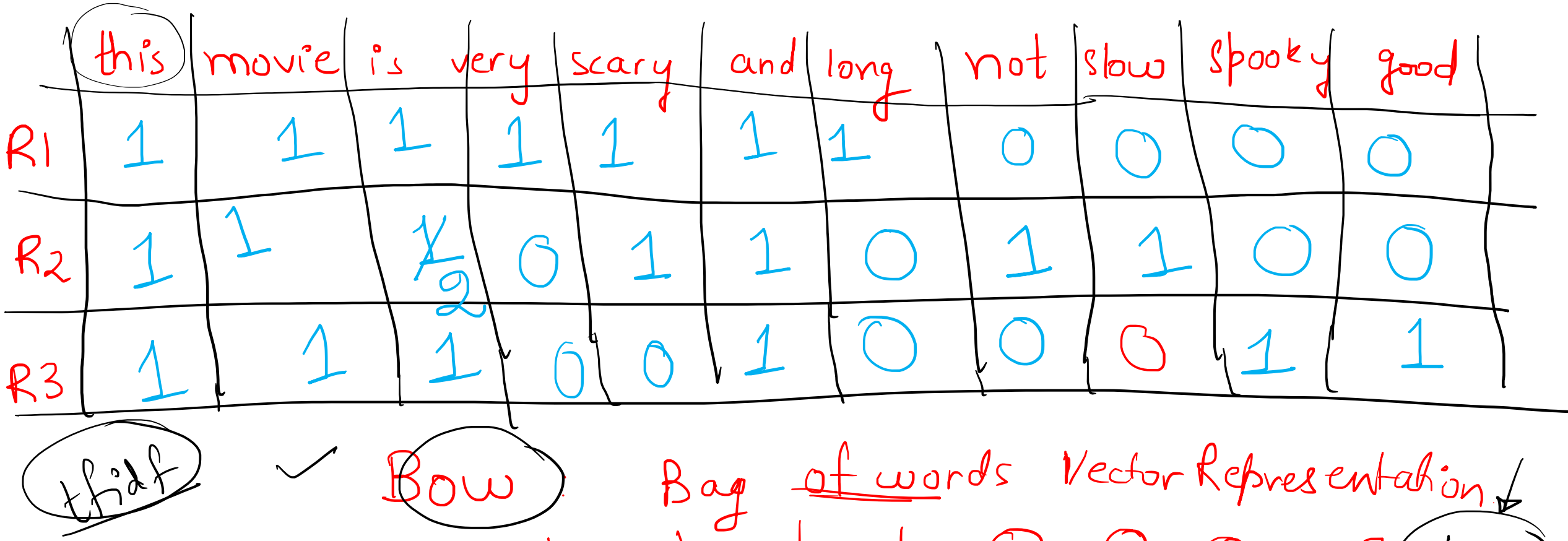
R2: This movie is not scary and is slow.

R3: This movie is spooky and good.

Bag of words :

Build a vocabulary = unique words from all doc.

11



tfidf ✓ BOW Bag of words Vector Representation

sparse

R1: [1 1 1 1 1 1 1 0 0 0 0]

R2: [1 1 2 0 1 1 0 1 1 0 0]

R3: [1 1 1 0 0 1 0 0 0 1 1]

⊛ Jaccard similarity $(A, B) = \frac{|A \cap B|}{|A \cup B|}$

** Dice coefficient $= \frac{2\left(|A \cap B|\right)}{|A| + |B|}$

|A| =

A S1: the cat sat on the mat

B S2: the kitten rested on the rug

the, cat, sat, on, the, mat, kitten, rested, rug |B| =

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{8} = \frac{1}{4} = 0.25$$

$\frac{1}{3} = 0.33$