

Corpus D

+1 →

d₁

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

1

2

3

4

5

6

7

8

9

10

11

12

TF-IDF = ("fox", d₁, D)

+1 →

d₂

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

1

2

3

4

5

6

7

8

9

10

11

12

TF-IDF = ("fox", d₂, D)

Question: How word **fox** is relevant to corpus D documents?

Solution:

TF is the frequency of any "term" in a given "document".

TF-IDF

IDF is constant per corpus, and accounts for the ratio of documents that include that specific "term".

$$TF(\text{"fox"}, d1) = 2 / 12 = 0.17$$

$$TF(\text{"fox"}, d2) = 3 / 12 = 0.25$$

$$IDF(\text{"fox"}, D) = \log(2/2) = 0$$