TF–IDF: Numerical Explanation and Information-Theoretic Derivation

Dr. Sambit Praharaj, Assistant Professor (II), KIIT

1 Introduction

TF–IDF (Term Frequency–Inverse Document Frequency) is a statistical weighting technique used in Natural Language Processing and Information Retrieval to measure the importance of a word in a document relative to a corpus.

$$\text{TF–IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

2 Example Corpus

Consider the following corpus consisting of three documents:

- D1: I love NLP
- D2: I love AI
- D3: I love AI and NLP

Total number of documents: $N = 3$

Vocabulary

$$\{I,\ love,\ AI,\ NLP,\ and\}$$

3 Term Frequency (TF)

Term Frequency measures the importance of a term within a document. It is necessary to normalize it so that relative importance of a term is considered proportional to the size of the document.

$$\text{TF}(t, d) = \frac{f(t, d)}{\sum_k f(k, d)}$$

TF Values

Word	TF in D1	TF in D2	TF in D3
I	1/3	1/3	1/5
love	1/3	1/3	1/5
AI	0	1/3	1/5
NLP	1/3	0	1/5
and	0	0	1/5

4 Inverse Document Frequency (IDF)

Document Frequency (DF) is the number of documents containing a term. N is the total number of documents.

$$\text{IDF}(t) = \log\left(\frac{N}{df(t)}\right)$$

DF and IDF Values

Word	DF	IDF
I	3	0
love	3	0
AI	2	0.176
NLP	2	0.176
and	1	0.477

5 TF–IDF Computation

TF–IDF is computed as the product of TF and IDF.

$$\text{TF–IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

TF–IDF for Document D1

Word	TF	IDF	TF–IDF
I	1/3	0	0
love	1/3	0	0
NLP	1/3	0.176	0.058
AI	0	0.176	0
and	0	0.477	0

TF–IDF for Document D2

Word	TF	IDF	TF–IDF
I	1/3	0	0
love	1/3	0	0
AI	1/3	0.176	0.058
NLP	0	0.176	0
and	0	0.477	0

TF–IDF for Document D3

Word	TF	IDF	TF–IDF
I	1/5	0	0
love	1/5	0	0
AI	1/5	0.176	0.035
NLP	1/5	0.176	0.035
and	1/5	0.477	0.095

6 Derivation of TF–IDF from Information Theory (NOT IN SYLLABUS BUT FOR UNDERSTANDING TF-IDF)

According to Shannon's Information Theory, the information content of an event x is:

$$I(x) = -\log P(x)$$

Information Content of a Term

The probability that a randomly chosen document contains term t is approximated as:

$$P(t) = \frac{df(t)}{N}$$

Thus, the information content of term t is:

$$I(t) = -\log\left(\frac{df(t)}{N}\right) = \log\left(\frac{N}{df(t)}\right)$$

This directly gives the IDF term.

Local Evidence in a Document

Term Frequency represents the probability of observing term t in document d:

$$\mathrm{TF}(t, d) = P(t \mid d)$$

Expected Information Contribution

The importance of a term is modeled as the expected information contribution:

$$\text{TF–IDF}(t, d) = P(t \mid d) \times I(t)$$

$$\boxed{\text{TF–IDF}(t, d) = \frac{f(t, d)}{\sum_k f(k, d)} \times \log\left(\frac{N}{df(t)}\right)}$$

7 Key Observations

- Words appearing in all documents have zero information content.
- Rare words have higher information value.
- TF–IDF balances local importance and global rarity.
- TF–IDF converts text into numerical vectors suitable for machine learning.

8 Conclusion

TF–IDF is a principled term weighting scheme derived from information theory, where IDF represents information content and TF represents evidence within a document. Its simplicity and interpretability make it a foundational technique in text mining and information retrieval.