TF-IDF: Term Frequency-Inverse Document Frequency

Your Name

March 24, 2025

Conceptual Overview

- TF-IDF aims to reflect how important a word is to a document in a collection (corpus).
- It combines two metrics:
  - **Term Frequency (TF)**: How often a term appears in a document.
  - **Inverse Document Frequency (IDF)**: How rare a term is across the entire corpus.

Example Corpus

Consider a small corpus of three call transcripts:

- Document 1 (D1): "Congratulations! You've won a prize. Please call to claim your reward."
- Document 2 (D2): "Urgent notice! Your account has been compromised. Call immediately to secure it."
- Document 3 (D3): "Hello, this is a standard call for account verification. Please verify your information."

Step 1: Term Frequency (TF)

TF measures how often each term appears in each document (raw count):

Term	D1	D2	D3
Congratulations	1	0	0
Won	1	0	0
Prize	1	0	0
Please	1	0	1
Call	1	1	1
Account	0	1	1
Secure	0	1	0
Verification	0	0	1

Table: Raw Term Frequency Count in Each Document

Step 2: Compute Term Frequency (TF)

Formula:

$$TF(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in } d}$$

Example for "Call" in D1 (assuming D1 has 10 words):

$$TF(\text{Call}, D1) = \frac{1}{10} = 0.1$$

Step 3: Compute Inverse Document Frequency (IDF)

Formula:

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

where:

- N = Total number of documents in the corpus.
- $df(t)$ = Number of documents containing term t.

Example for "Call":

$$IDF(\text{Call}) = \log\left(\frac{3}{3}\right) = \log(1) = 0$$

Example for "Congratulations":

$$IDF(\text{Congratulations}) = \log\left(\frac{3}{1}\right) = \log(3) \approx 0.477$$

Step 4: Compute TF-IDF

Formula:

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

Example for "Congratulations" in D1:

$$TF\text{-}IDF(\text{Congratulations}, D1) = 0.1 \times 0.477 = 0.0477$$

Example for "Call" (appears in all documents, IDF = 0):

$$TF\text{-}IDF(\text{Call}, D1) = 0.1 \times 0 = 0$$

Final TF-IDF Table

Term	D1	D2	D3
Congratulations	0.0477	0	0
Won	0.0477	0	0
Prize	0.0477	0	0
Please	0.0477	0	0.0477
Call	0	0	0
Account	0	0.0477	0.0477
Secure	0	0.477	0
Verification	0	0	0.477

Table: TF-IDF Scores for Terms in Each Document

Conclusion

- TF-IDF helps identify important terms in a document.
- Words common in all documents (like "Call") get **low** TF-IDF scores.
- Rare words (like "Secure" and "Verification") get **high** TF-IDF scores, highlighting their importance.
- This numerical representation of text is then used as features for machine learning models.