#1 Term Frequency

Fundamental concepts for understanding TF-IDF

Key Definitions:

- **Corpus (D)**: A collection of N documents

  $D = \{d_i\}_{i=1}^{N}$ where $i \in \{1, 2, ..., N\}$

- **Vocabulary (V)**: All unique terms across the corpus

  $V = \{t_1, t_2, ..., t_{|V|}\}$ where $|V|$ is the vocabulary size

- **Term Frequency**: How often term t appears in document d_i

  Raw TF: $\mathrm{TF}_{\mathrm{raw}}(t, d_i) = f_{t,d_i}$

  Normalized TF: $\mathrm{TF}_{\mathrm{norm}}(t, d_i) = \frac{f_{t,d_i}}{\sum_{t' \in d_i} f_{t',d_i}}$

  Log-scaled TF: $\mathrm{TF}_{\mathrm{log}}(t, d_i) = \log(1 + f_{t,d_i})$

@AIinMinutes

#2 Document Frequency

Capturing the rarity and importance of terms across documents

Formula:

$$\mathrm{DF}(t) = |\{d_i \in D : t \in d_i\}|$$

Terms Explanation:

- $\mathrm{DF}(t)$: Number of documents containing term t

- $\{d_i \in D : t \in d_i\}$: Set of documents containing t

- High DF indicates common terms across corpus (e.g., "the", "is")

Key Insight: If a term is common across the corpus, its high term frequency in a particular document doesn't reveal any characteristic of that document

@AIinMinutes

#3 Inverse Document Frequency (IDF)

Quantifying the importance of rare terms across the corpus

Formula:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{|\{d_i \in D : t \in d_i\}|}\right)$$

Terms Explanation:

- N: Total number of documents in corpus

- $\mathrm{DF}(t) = |\{d_i \in D : t \in d_i\}|$: Document frequency

- Log scaling: Accounts for Zipf's law (power law distribution of terms)

Smoothed IDF:

$$\mathrm{IDF}_{\mathrm{smooth}}(t) = \log\left(\frac{N+1}{|\{d_i \in D : t \in d_i\}|+1}\right)$$

@AIinMinutes

#4 TF-IDF Calculation

Combining term frequency and inverse document frequency

Formula:

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

Terms Explanation:

- $\text{TF}(t, d_i)$: Term frequency of t in document d_i

- $\text{IDF}(t)$: Inverse document frequency of term t

- High TF-IDF: Term appears frequently in document but rarely in corpus

$\Longrightarrow$ term t is representative of the document d_i

@AIinMinutes