Tutorial

Calculating TF-IDF For Three Documents

Course:

Natural Language Processing

By:

Dr. Zohair Ahmed

Example Corpus

1. **Document 1**: "Machine learning is fun."
2. **Document 2**: "Deep learning is fascinating."
3. **Document 3**: "Deep learning is the future."

Step 1: Calculate Term Frequency (TF)

Term Frequency (TF) is the frequency of a word in a document, defined as:

$$\text{TF(w)} = \frac{\text{Number of times word w appears in a document}}{\text{Total number of words in the document}}$$

Document 1: "Machine learning is fun."

- Words: ["Machine", "learning", "is", "fun"]
- TF for each word:
  - TF("Machine") = 1/4 = **0.25**
  - TF("learning") = 1/4 = **0.25**
  - TF("is") = 1/4 = **0.25**
  - TF("fun") = 1/4 = **0.25**

Document 2: "Deep learning is fascinating."

- Words: ["Deep", "learning", "is", "fascinating"]
- TF for each word:
  - TF("Deep") = 1/4 = **0.25**
  - TF("learning") = 1/4 = **0.25**
  - TF("is") = 1/4 = **0.25**
  - TF("fascinating") = 1/4 = **0.25**

Document 3: "Deep learning is the future."

- Words: ["Deep", "learning", "is", "the", "future"]
- TF for each word:
  - TF("Deep") = 1/5 = **0.20**
  - TF("learning") = 1/5 = **0.20**
  - TF("is") = 1/5 = **0.20**
  - TF("the") = 1/5 = **0.20**
  - TF("future") = 1/5 = **0.20**

Step 2: Calculate Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) is defined as:

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

Where:

- **N** = total number of documents.
- **df(w)** = number of documents that contain the word **w**.

Document Frequency (df) Calculation:

- **"Machine"** appears in Document 1 → **df("Machine") = 1**.
- **"learning"** appears in all three documents → **df("learning") = 3**.
- **"is"** appears in all three documents → **df("is") = 3**.
- **"fun"** appears in Document 1 → **df("fun") = 1**.
- **"Deep"** appears in Documents 2 and 3 → **df("Deep") = 2**.
- **"fascinating"** appears in Document 2 → **df("fascinating") = 1**.
- **"the"** appears in Document 3 → **df("the") = 1**.
- **"future"** appears in Document 3 → **df("future") = 1**.

IDF Calculation:

Let's calculate the IDF for some words:

- **IDF("Machine")**:

$$\text{IDF(Machine)} = \log\left(\frac{3}{1}\right) = \log(3) \approx 1.1$$

- **IDF("learning")**

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

(Since it appears in all documents, it gets a low IDF, reflecting that it is not a rare word.)

- **IDF("Deep")**:

$$\text{IDF(Deep)} = \log\left(\frac{3}{2}\right) \approx 0.18$$

- **IDF("fascinating")**:

$$\text{IDF(fascinating)} = \log\left(\frac{3}{1}\right) = \log(3) \approx 1.1$$

Step 3: Calculate TF-IDF

Now, let's calculate the **TF-IDF** for some words using the formula:

$$\text{TF-IDF}(w) = \text{TF}(w) \times \text{IDF}(w)$$

TF-IDF for "Machine" in Document 1:

- TF("Machine") = 0.25
- IDF("Machine") = 1.1

$$\text{TF-IDF}(\text{Machine}) = 0.25 \times 1.1 = 0.275$$

TF-IDF for "learning" in Document 1:

- TF("learning") = 0.25
- IDF("learning") = 0

$$\text{TF-IDF}(\text{learning}) = 0.25 \times 0 = 0$$

TF-IDF for "Deep" in Document 2:

- TF("Deep") = 0.25
- IDF("Deep") = 0.18

$$\text{TF-IDF}(\text{Deep}) = 0.25 \times 0.18 = 0.045$$

TF-IDF for "fascinating" in Document 2:

- TF("fascinating") = 0.25
- IDF("fascinating") = 1.1

$$\text{TF-IDF}(\text{fascinating}) = 0.25 \times 1.1 = 0.275$$

Final Results: TF-IDF for Selected Words

Word	Document 1 (TF-IDF)	Document 2 (TF-IDF)	Document 3 (TF-IDF)
Machine	0.275	0	0
learning	0	0	0
is	0.25	0.25	0.25
fun	0.25	0	0
Deep	0	0.045	0.045
fascinating	0	0.275	0
the	0	0	0.2

future	0	0	0.2

Most Important

- **Rare words** like "Machine" and "fascinating" have **higher TF-IDF** because they are rare in the corpus, despite their lower frequency in the document.
- **Common words** like "learning" and "is" have **low TF-IDF** scores because they appear in all documents, reducing their importance.
- **TF-IDF** highlights the **importance of rare words** in a specific document while considering their rarity in the entire corpus.

How TF and IDF Control

- **TF** is the **frequency of a word in a document**, not across all documents.
- **IDF (Inverse Document Frequency)** measures how important a word is across all documents in the corpus.
    - **IDF** increases when a word is rare across the corpus and decreases when the word is common in many documents. It helps **down-weight common words** and **emphasize rare words**.
    - **The value of IDF decreases** when the word is present in more documents, and it increases when the word is rare (appears in fewer documents). The **log** just controls the scaling.
    - **Higher denominator** (more documents with the word) leads to a **lower IDF**
    - **Lower denominator** (fewer documents with the word) leads to a **higher IDF**