Problem 1: Calculating Term Frequency (TF)

Problem:
Given a document with the following text:
"Text mining is a process of deriving high-quality information from text."

Calculate the term frequency (TF) for the term "text" in the document.

Solution:

1. **Term Frequency (TF)**: It is the number of times a term appears in a document divided by the total number of terms in the document.
   - **Count of term "text"**: 2 (appears twice in the document)
   - **Total number of terms in the document**: 9

$$\text{TF}(\text{"text"}) = \frac{\text{Count of "text"}}{\text{Total number of terms}} = \frac{2}{9} \approx 0.222$$

   So, the TF for "text" is approxi[illegible]tely 0.222.

Problem 2: Calculating Inverse Document Frequency (IDF)

Problem:
You have a corpus of 100 documents. The term "mining" appears in 10 of these documents. Calculate the inverse document frequency (IDF) for the term "mining".

Solution:

1. **Inverse Document Frequency (IDF)**: It is calculated as:

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

   where:
   - N is the total number of documents (100 in this case).
   - $df(t)$ is the number of documents containing the term (10 in this case).

$$\text{IDF}(\text{"mining"}) = \log\left(\frac{100}{10}\right) = \log(10) = 1$$

So, the IDF for "mining" is 1.

Problem 3: Calculating TF-IDF

Problem:
Given a document where the term "text" appears 3 times out of 10 total terms, and assuming the IDF for "text" is 0.5, calculate the TF-IDF score for the term "text".

Solution:

1. **TF-IDF Score**: It is calculated as:

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

where:

- **TF** is calculated as $\frac{\text{Count of term}}{\text{Total terms in document}}$.

$$\text{TF}(\text{"text"}) = \frac{3}{10} = 0.3$$

$$\text{TF-IDF}(\text{"text"}) = 0.3 \times 0.5 = 0.15$$

So, the TF-IDF score for "text" is 0.15.

Problem 4: Cosine Similarity between Two Documents

Problem:
Calculate the cosine similarity between the following two vectors representing term frequencies in two documents:

- Document 1: $[2, 3, 0, 5]$
- Document 2: $[1, 0, 4, 2]$

Solution:

1. **Cosine Similarity**: It is calculated as:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where:

- $\mathbf{A}$ and $\mathbf{B}$ are the term frequency vectors for the two documents.
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes of the vectors.

$$\mathbf{A} \cdot \mathbf{B} = (2 \times 1) + (3 \times 0) + (0 \times 4) + (5 \times 2) = 2 + 0 + 0 + 10 = 1$$

$$\|\mathbf{A}\| = \sqrt{(2^2) + (3^2) + (0^2) + (5^2)} = \sqrt{4 + 9 + 0 + 25} = \sqrt{38} \approx 6.1$$

$$\|\mathbf{B}\| = \sqrt{(1^2) + (0^2) + (4^2) + (2^2)} = \sqrt{1 + 0 + 16 + 4} = \sqrt{21} \approx 4.5$$

$$\text{Cosine Similarity} = \frac{12}{6.16 \times 4.58} \approx \frac{12}{28.21} \approx 0.426$$

So, the cosine similarity between the two documents is approximately 0.426.

Problem 5: K-Means Clustering (Centroid Calculation)

Problem:

Given the following data points and initial centroids for K-Means clustering:

- Data points: $(1, 2)$, $(2, 3)$, $(3, 4)$, $(10, 10)$, $(11, 11)$
- Initial centroids: $(1, 2)$ and $(10, 10)$

Calculate the new centroids after one iteration.

Solution:

1. **Assign Points to Clusters**:
   - Cluster 1 (centroid $(1, 2)$): $(1, 2)$, $(2, 3)$, $(3, 4)$
   - Cluster 2 (centroid $(10, 10)$): $(10, 10)$, $(11, 11)$
2. **Calculate New Centroids**:

- **Cluster 1**:

$$\text{New Centroid}_1 = \left(\frac{1+2+3}{3}, \frac{2+3+4}{3}\right) = \left(\frac{6}{3}, \frac{9}{3}\right) = (2, 3)$$

- **Cluster 2**:

$$\text{New Centroid}_2 = \left(\frac{10+11}{2}, \frac{10+11}{2}\right) = \left(\frac{21}{2}, \frac{21}{2}\right) = (10.5, 1$$

So, the new centroids are $(2, 3)$ and $(10.5, 10.5)$.

These problems and solutions illustrate key numerical calculations in text mining, including term frequency, inverse document frequency, TF-IDF, cosine similarity, and K-Means clustering.