Top 20 Interview Questions on TF-IDF (with Answers)

1. What is TF-IDF?

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate how important a word is to a document in a collection. It combines two components: TF (how often a term appears in a document) and IDF (how rare the term is across all documents).

2. Why use TF-IDF instead of raw term frequency?

Raw term frequency may overvalue common words. TF-IDF penalizes frequent terms and highlights rare, informative terms, improving feature quality.

3. What does Term Frequency (TF) mean?

TF measures how often a word appears in a document. It's usually normalized by the total word count.

4. What does Inverse Document Frequency (IDF) mean?

IDF measures how unique or rare a word is across all documents. Words that appear in many documents get lower IDF scores.

5. How is TF-IDF calculated?

TF-IDF = TF * IDF. TF is term count divided by total terms in the document. IDF = log(N / df), where N is total documents and df is documents containing the term.

6. What is the effect of smoothing in IDF?

Smoothing avoids zero or infinite IDF values. It's done by adding 1 to numerator and denominator: IDF = log((1+N)/(1+df)) + 1.

7. Why is log-scaling used in IDF?

Log-scaling compresses large IDF values, preventing rare words from dominating the importance score.

8. What is Sublinear TF scaling?

It dampens high term frequency by using TF = 1 + log(tf) instead of raw counts. This avoids biasing scores for terms that repeat too much.

9. Is sublinear TF scaling the same as smoothing?

No. Sublinear scaling adjusts term frequency within a document. Smoothing adjusts IDF to handle rare or common words across documents.

10. What does a low TF-IDF score mean?

The term is either common across documents (high df) or appears infrequently in a specific document (low tf), meaning it carries less importance.

11. Why is TF-IDF considered unsupervised?

It doesn't require labels or predefined categories. TF-IDF is computed purely from term and document statistics.

12. How does document length affect TF-IDF?

Longer documents can have higher TF, so TF is often normalized. This ensures longer docs don't overpower shorter ones unfairly.

13. How is TF-IDF used in clustering?

TF-IDF vectors represent documents numerically. Algorithms like K-Means use them to group similar documents based on their content.

14. How does corpus size affect TF-IDF?

As corpus size grows, IDF values adjust. A term becomes more rare (higher IDF) or more common (lower IDF), changing its TF-IDF score.

15. What is a TF-IDF matrix?

A 2D matrix where rows are documents, columns are terms (features), and each cell contains the TF-IDF score of a term in a document.

16. In the TF-IDF matrix, what is a feature?

Each column (term/word) is a feature. These features are used to represent text numerically in ML models.

17. How do you select top features from a TF-IDF matrix?

By using max TF-IDF per term, average TF-IDF, or supervised methods like Chi-square or Mutual Information.

18. How does maximum TF-IDF help in feature selection?

It selects terms that are highly important in at least one document, capturing uniqueness.

19. How does average TF-IDF help in feature selection?

It selects terms that are consistently important across many documents, ensuring broad relevance.

20. How do Chi-square and Mutual Information select features?

These supervised methods score terms by how strongly they relate to class labels, useful in classification tasks.