Top 20 Bag of Words (BoW) Interview Questions with Answers

1. What is the Bag of Words (BoW) model in NLP? [Basic]

BoW is a text representation method that treats a document as a collection ("bag") of its words. It ignores grammar and word order, focusing only on word frequency or presence.

2. What are the steps to build a BoW model? [Basic]

1. Tokenize text
2. Create vocabulary
3. Count word occurrences
4. Represent each document as a frequency vector

3. What are the advantages of using BoW? [Basic]

- Simple and easy to implement
- Effective for small text tasks
- Useful for document classification and spam detection

4. What are the limitations of BoW? [Basic]

- Ignores word order and context
- High dimensionality
- Cannot capture semantics or similarity in meaning

5. How does BoW handle stop words? [Basic]

Stop words (e.g., "the", "and", "is") are often removed during preprocessing to reduce noise and dimensionality before constructing the BoW.

6. What is the difference between frequency-based BoW and binary BoW? [Medium]

- Frequency-based: Counts how many times each word appears
- Binary BoW: Just checks if a word exists (1) or not (0)

7. How does vocabulary size impact a BoW model? [Medium]

A larger vocabulary increases feature dimensions, which leads to more sparse vectors, higher memory usage, and potentially overfitting.

8. Why is BoW considered a sparse representation? [Medium]

Most documents use only a small subset of the full vocabulary, so most entries in the vector are zeros-making it sparse.

9. How do n-grams improve BoW representation? Give an example. [Medium]

N-grams capture word sequences. For example, bi-grams can distinguish "not good" from "very good", which unigram BoW would miss.

10. What preprocessing is done before applying BoW? [Medium]

- Lowercasing
- Tokenization
- Stop word removal
- Stemming/Lemmatization
- Removing punctuation/special characters

11. Can two different sentences have the same BoW representation? Give an example. [Hard]

Yes. Example: "Dogs chase cats" vs "Cats chase dogs" -> same BoW, but meaning differs due to ignored word order.

12. Why is BoW unsuitable for tasks like machine translation or question answering? [Hard]

BoW ignores word order and context, making it incapable of understanding syntax or semantics, both critical in such tasks.

13. What are the mathematical limitations of BoW? [Hard]

BoW cannot model semantics, context, or word similarity. It treats each word as an independent feature and

lacks vector space relationships.

14. What is the role of dimensionality reduction in BoW? [Hard]

It reduces sparsity and overfitting, improves computation, and keeps only the most informative features using PCA, LSA, or feature selection.

15. How does TF-IDF improve upon the basic BoW model? [Hard]

TF-IDF reduces the impact of frequently used words (like "the") and gives higher importance to rare but informative words.

16. How can feature selection methods improve BoW-based models? [Hard]

Methods like Chi-square or mutual information remove irrelevant or redundant words, enhancing accuracy and reducing overfitting.

17. Compare BoW and Word2Vec in terms of context and representation. [Hard]

- BoW: Sparse, no context
- Word2Vec: Dense, contextual

Word2Vec captures semantic relationships and meaning; BoW does not.

18. How would you use BoW in a spam classification system? [Hard]

Preprocess -> Convert emails to BoW vectors -> Train classifier (Naive Bayes/SVM) -> Predict if new email is spam or not.

19. How can BoW features be combined with other metadata features? [Hard]

BoW can be concatenated with features like email sender, time, user data, etc., to form a richer feature set for better model performance.

20. How do you scale BoW for large corpora with millions of documents? [Hard]

Use:

- Hashing trick (HashVectorizer)
- Limit vocabulary size (top-K frequent words)
- Sparse matrix representations
- Distributed storage and processing