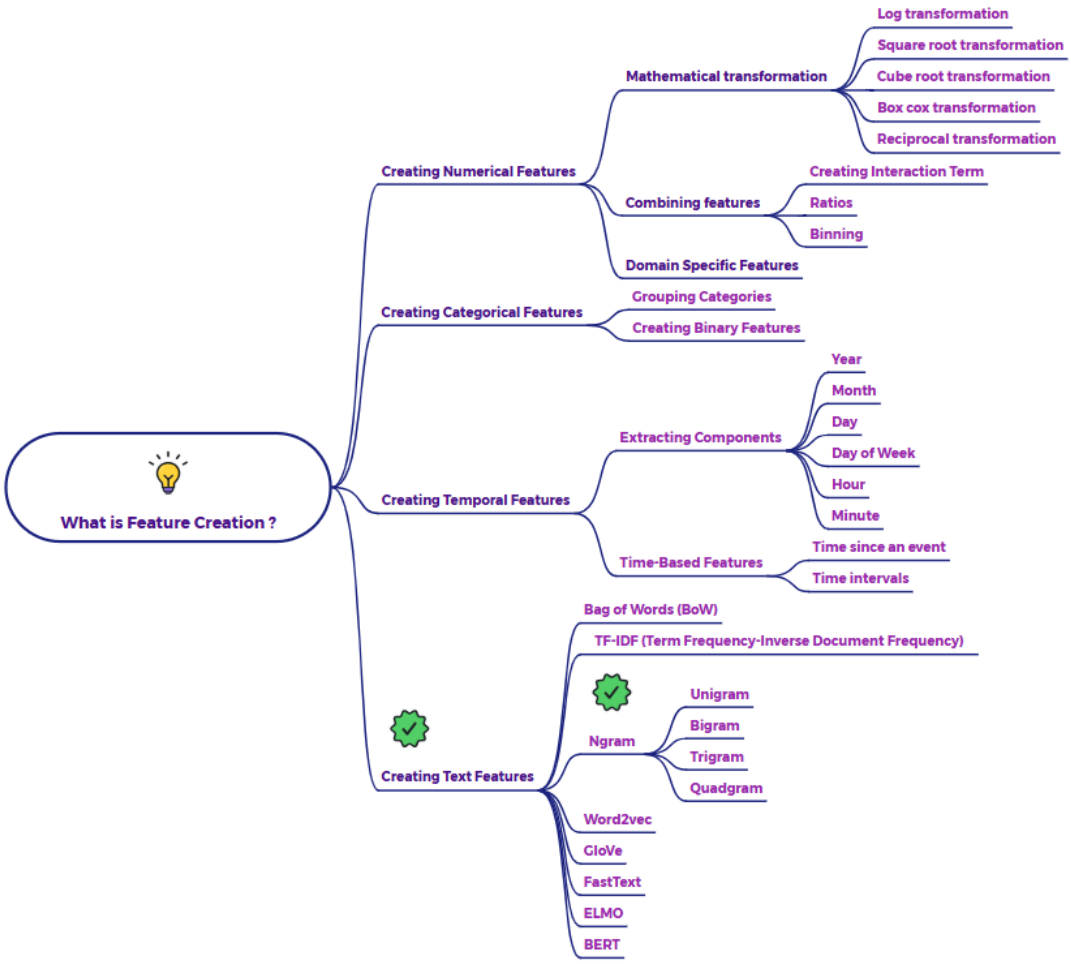
Explain N-gram with an example



N-grams

N-grams are contiguous sequences of n items from a given sequence of text or speech. In natural language processing (NLP), they are used to capture the local order of words in a sentence. The "n" in n-gram refers to the number of words in the sequence.

Here's a breakdown with examples:

1. Unigram

- A unigram is a sequence of one word.
- Example:
    - Sentence: "The quick brown fox"
    - Unigrams: "The", "quick", "brown", "fox"

2. Bigram

- A bigram is a sequence of two adjacent words.
- Example:
  - Sentence: "The quick brown fox"
  - Bigrams: "The quick", "quick brown", "brown fox"

3. Trigram

- A trigram is a sequence of three adjacent words.
- Example:
  - Sentence: "The quick brown fox"
  - Trigrams: "The quick brown", "quick brown fox"

4. Quadgram

- A quadgram is a sequence of four adjacent words.
- Example:
  - Sentence: "The quick brown fox"
  - Quadgram: "The quick brown fox"

Example with a longer sentence:

- Sentence: "I went to the beautiful park and saw many colorful flowers"
- Unigrams: "I", "went", "to", "the", "beautiful", "park", "and", "saw", "many", "colorful", "flowers"
- Bigrams: "I went", "went to", "to the", "the beautiful", "beautiful park", "park and", "and saw", "saw many", "many colorful", "colorful flowers"
- Trigrams: "I went to", "went to the", "to the beautiful", "the beautiful park", "beautiful park and", "park and saw", "and saw many", "saw many colorful", "many colorful flowers"
- Quadgrams: "I went to the", "went to the beautiful", "to the beautiful park", "the beautiful park and", "beautiful park and saw", "park and saw many", "and saw many colorful", "saw many colorful flowers"

Key Points

- N-grams capture some context within a sentence. For example, "quick brown" provides more information than "quick" or "brown" alone.
- As 'n' increases, the n-grams capture more context, but the number of possible n-grams also increases rapidly, leading to higher dimensionality.
- N-grams are used in various NLP tasks, including language modeling, text classification, and information retrieval.