Lab 1

Bag of Words (BoW) Model: Formula and Explanation

The Bag of Words model represents a text document as a numerical feature vector, where:

1. Each unique word in the document corpus is treated as a feature.
2. The value of a feature is the frequency of the corresponding word in the document.

Steps in the BoW Model

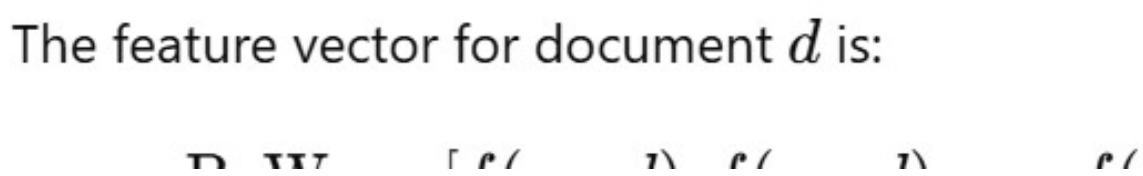
1. **Tokenization:** Break text into individual words.
2. **Build Vocabulary:** Create a list of all unique words in the corpus.
3. **Vectorization:** Count the frequency of each word in the vocabulary for every document.

Mathematical Representation

Let:

- D: Total number of documents.
- V: Vocabulary (set of unique words across all documents).
- f(w,d): Frequency of word w in document d.

The feature vector for document d is:



Example of Bag of Words

Corpus of Documents:

1. **Document 1:** "Text mining is fun."
2. **Document 2:** "Text mining helps extract insights."

Step 1: Tokenization

- Document 1: `["text", "mining", "is", "fun"]`
- Document 2: `["text", "mining", "helps", "extract", "insights"]`
- **Step 2: Build Vocabulary**
- Unique words across both documents:

- V={"text", "mining", "is", "fun", "helps", "extract", "insights"}.V = \{ \text{"text", "mining", "is", "fun", "helps", "extract", "insights"} \}.V={"text", "mining", "is", "fun", "helps", "extract", "insights"}.

Step 3: Create Frequency Table

Word	Document 1 Frequency	Docum
text	1	1
mining	1	1
is	1	0
fun	1	0
helps	0	1

Step 4: Feature Vectors

- **Document 1 Vector:** `[1, 1, 1, 1, 0, 0, 0]`
- **Document 2 Vector:** `[1, 1, 0, 0, 1, 1, 1]`

Numerical Example for Machine Learning

Using the above feature vectors, you can input them into a machine learning model for classification, clustering, or other tasks.

For example, if the documents are labeled:

- Document 1: Positive Sentiment
- Document 2: Neutral Sentiment

The BoW vectors along with labels can be used to train a sentiment analysis classifier.