TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is a commonly used technique in natural language processing and information retrieval to convert a collection of text documents to a numerical format that can be used for machine learning algorithms. TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents (corpus). It consists of two main components:

1. Term Frequency (TF): It measures how often a term (word) appears in a document. It is calculated as the number of times a term occurs in a document divided by the total number of terms in that document. The idea is to give higher weight to words that appear more frequently in a document.

2. Inverse Document Frequency (IDF): It measures the importance of a term across the entire document corpus. Words that are common across many documents receive lower weights, while words that are rare or unique to a document receive higher weights.

Scikit-learn is a popular Python library that provides a `TfidfVectorizer` class to easily convert a collection of text documents to a TF-IDF representation. Here's a simple example:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
# Example documents
documents = ["This is the first document.",
             "This document is the second document.",
             "And this is the third one.",
             "Is this the first document?"]

# Create a TF-IDF vectorizer
vectorizer = TfidfVectorizer()

# Fit and transform the documents
tfidf_matrix = vectorizer.fit_transform(documents)

# Get the feature names (terms)
feature_names = vectorizer.get_feature_names_out()

# Convert the TF-IDF matrix to a dense array for better readability
dense_array = tfidf_matrix.toarray()

# Display the results
print("TF-IDF Matrix:")
print(dense_array)
print("\nFeature Names:")
print(feature_names)
```

Bag of Words

1. Bag of Words (BoW):

- Definition: Bag of Words is a commonly used technique in natural language processing where a text (such as a sentence or document) is represented as an unordered set of words, disregarding grammar and word order but keeping track of the frequency of each word.
- Process: The process involves creating a vocabulary of unique words from the entire set of documents (corpus) and then representing each document as a vector with the count (or presence/absence) of each word from the vocabulary.

Document 1: "I love natural language processing."
Document 2: "Natural language processing is fascinating."

Vocabulary: ["I", "love", "natural", "language", "processing", "is", "fascinating"]

BoW Representation:
Document 1: [1, 1, 1, 1, 1, 0, 0]
Document 2: [0, 0, 1, 1, 1, 1, 1]
```
```