LAB NO. 4

Date of Performance: 22-01-2025

**Aim:** Write Python Program to compute and display term Frequency inverse document frequency and bag of words representation for given text corpus write a python program to generate and display a word cloud from given text corpus.

## Objectives:

* **Computing and displaying Term Frequency-Inverse Document Frequency (TF-IDF):**  
  This helps quantify the importance of words in the corpus, balancing their frequency and uniqueness across documents.
* **Generating a Bag of Words (BoW) representation:**  
  This involves creating a numerical representation of the text corpus, providing insights into word occurrences without considering their order.
* **Creating and visualizing a Word Cloud:**  
  The word cloud will graphically represent the most frequent terms in the corpus, enabling a quick overview of key topics and themes.

**Resources Used:** Google Colab

## Date Of Performace : 22 / 02 / 2025

## Theory:

Text Representation in NLP

In Natural Language Processing (NLP), representing text numerically is critical for machine learning algorithms to process and analyze text data. Two popular methods for this are Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Additionally, Word Clouds are visual representations of text data, offering insights into the most frequent terms.

**1. Bag of Words (BoW)**

The Bag of Words model is a simple and widely used technique for representing text data. It transforms a text corpus into a numerical format by counting the occurrences of each word in the text.

Key Characteristics of BoW:

1. Text to Numbers: Each document is represented as a vector with the count of each word.
2. Vocabulary Creation: A dictionary of unique words (vocabulary) is created.
3. Order Ignored: It disregards the word order and focuses on the frequency of words.

Example:

For two sentences:

* Sentence 1: "I love coding."
* Sentence 2: "Coding is fun."

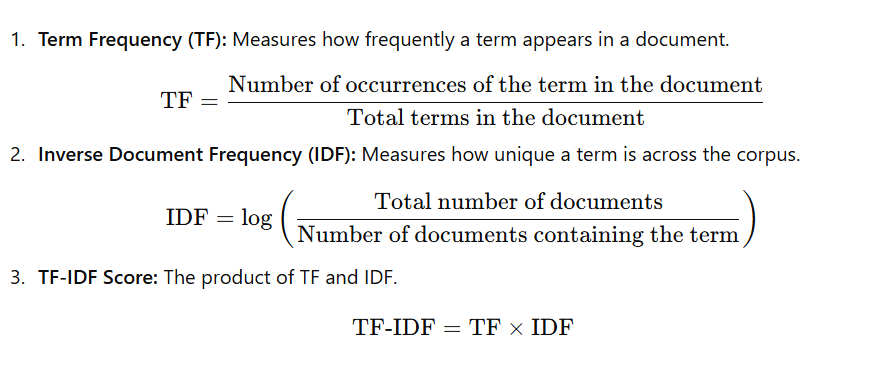
The vocabulary would be: ["I", "love", "coding", "is", "fun"].  
The BoW representation:

* Sentence 1: [1, 1, 1, 0, 0]
* Sentence 2: [0, 0, 1, 1, 1]

**2. Term Frequency-Inverse Document Frequency (TF-IDF)**

TF-IDF is an enhancement of the BoW model. It measures the importance of a word in a document relative to its frequency across the entire corpus.

Components of TF-IDF:



**3. Word Cloud**

A Word Cloud is a graphical representation of text data where the size of each word indicates its frequency or importance in the text. It is a useful visualization tool to gain insights into the most prominent terms in a corpus.

Applications of Word Clouds:

* Summarizing large text data.
* Visualizing frequently occurring terms in documents.
* Identifying key topics or themes.

Example of Word Cloud Usage:

For a document containing the sentence: "Data science is fun and useful," the Word Cloud may display "data" and "science" in larger fonts if they are repeated more frequently in the text.

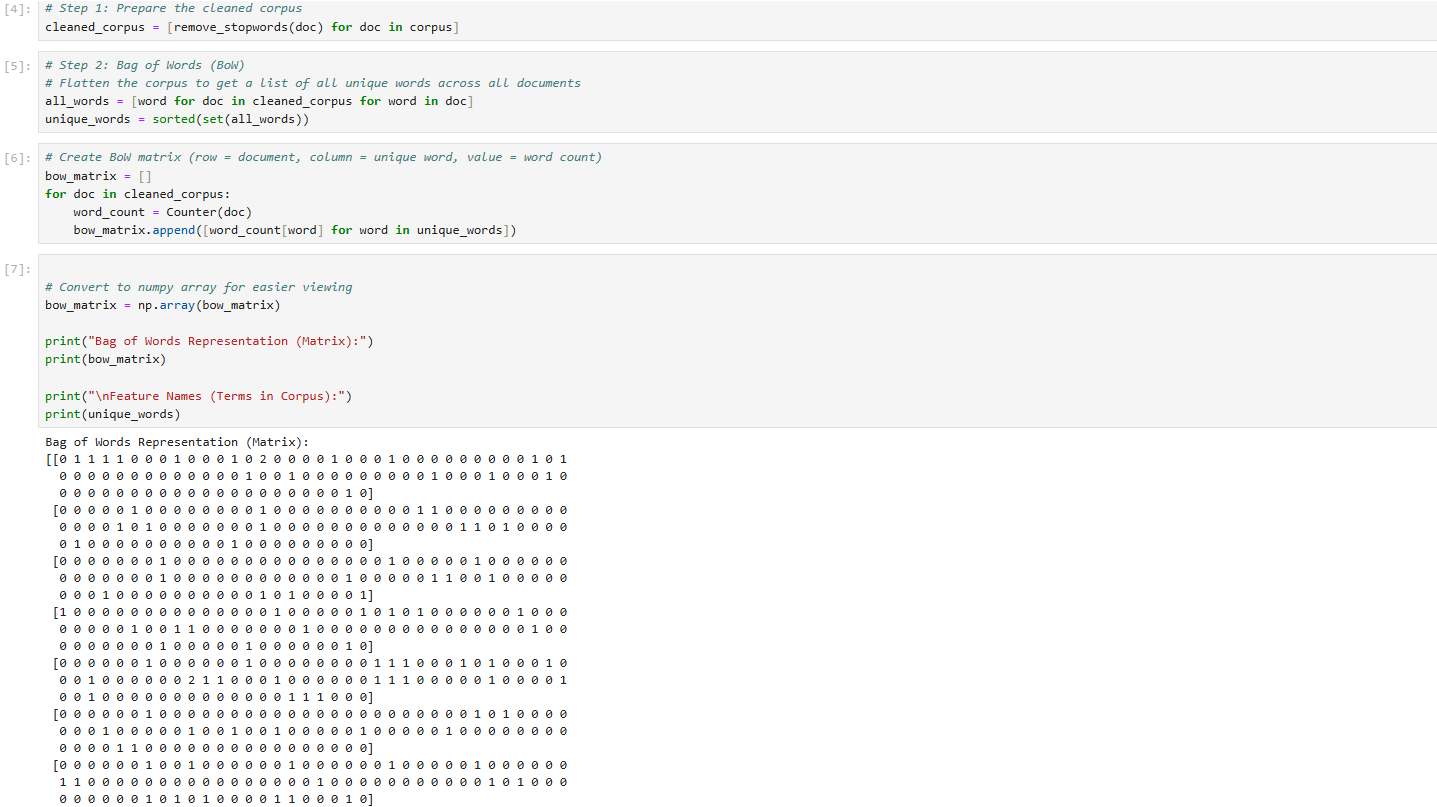
**Procedure Overview**

1. Input Data: Define the text corpus to be analyzed.
2. Bag of Words Representation: Use libraries like sklearn to generate the BoW matrix.
3. TF-IDF Computation: Compute the TF-IDF scores for words using TfidfVectorizer from sklearn.
4. Word Cloud Generation: Use the WordCloud library to create a visual representation of the text data.
5. Display Results: Print the BoW matrix, TF-IDF scores, and display the Word Cloud.

**Code /OutPut:**

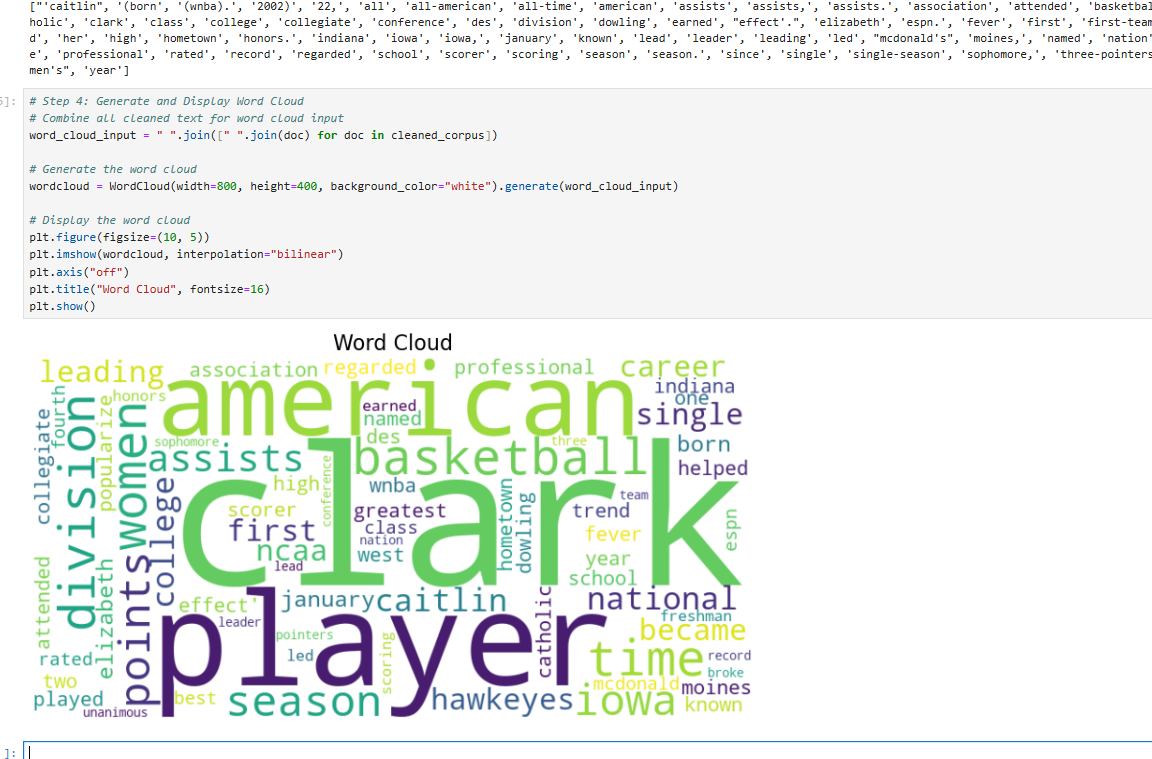
Without sklearn



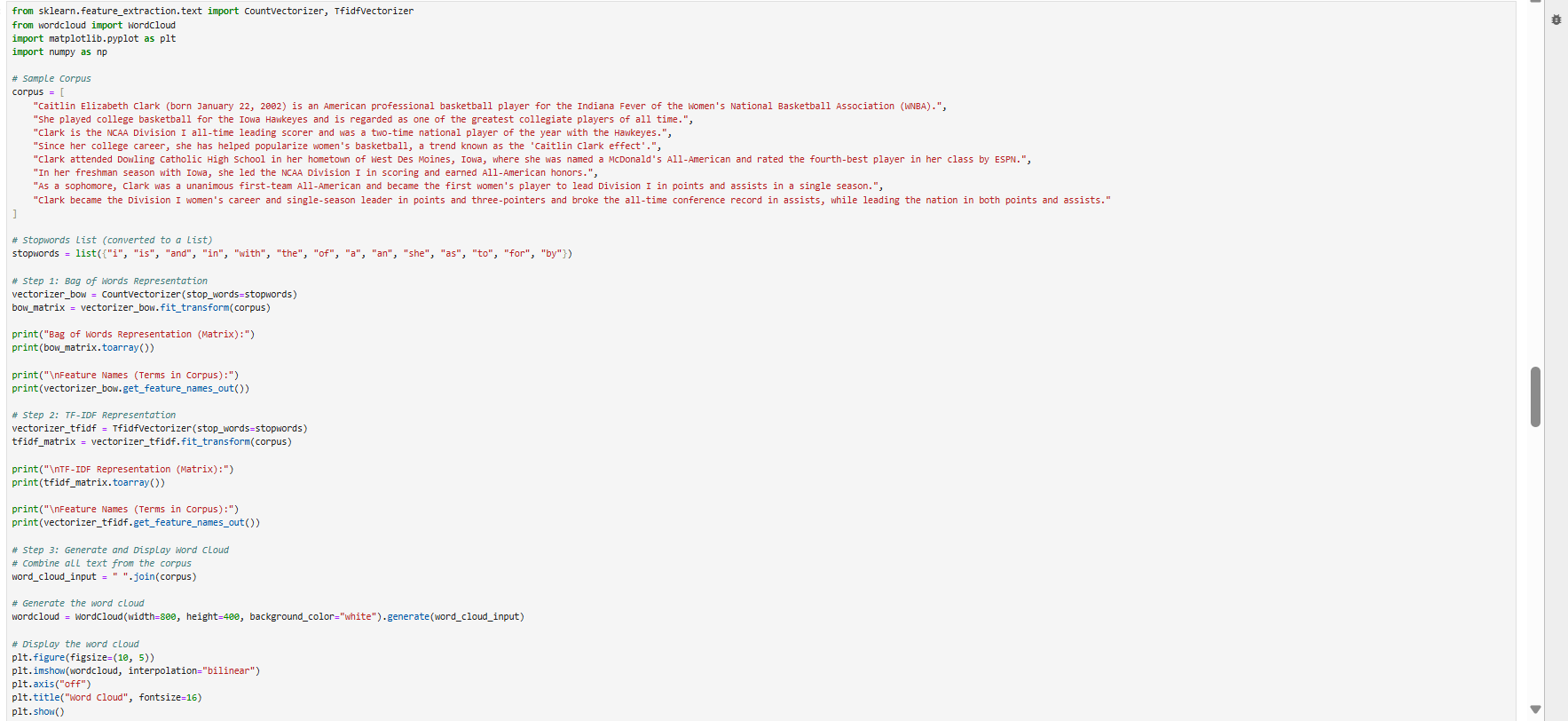


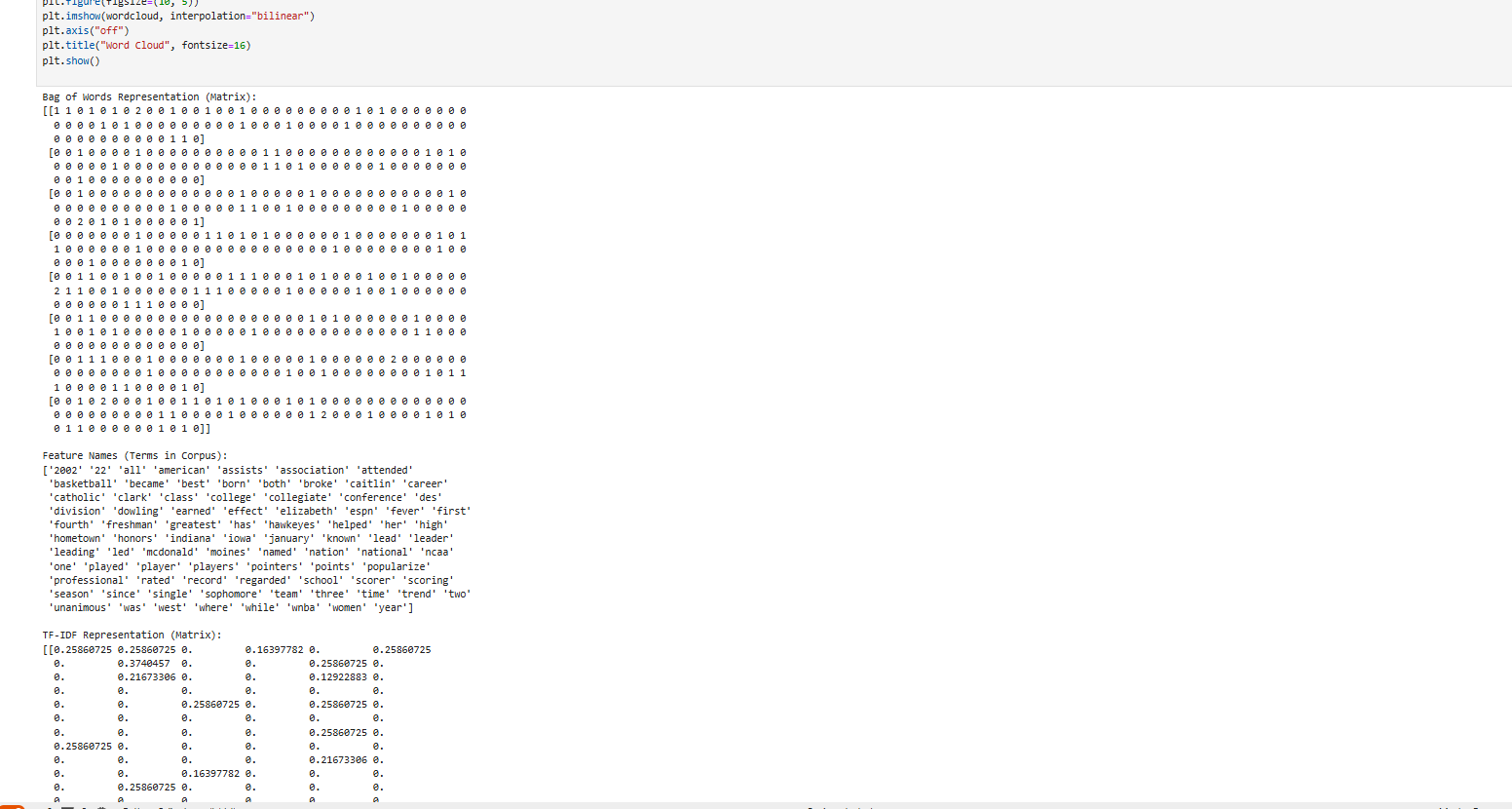


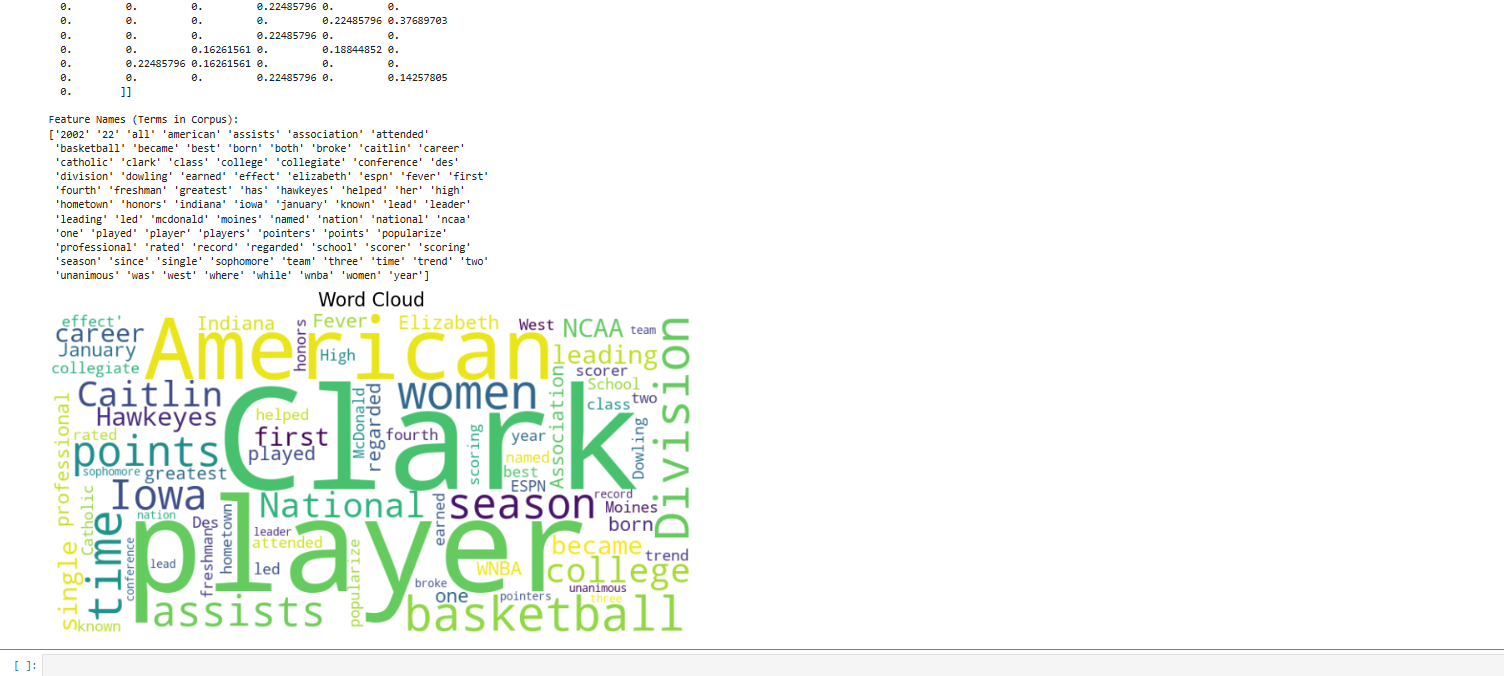




With Sklearn:







Viva Question :

**1. What is the purpose of the Term Frequency-Inverse Document Frequency (TF-IDF) technique?**

**Answer:** TF-IDF is used to evaluate the importance of a word in a document relative to its presence across the entire corpus. It highlights unique and relevant terms by assigning higher weights to terms that occur frequently in a document but are rare across the corpus.

**2. How does the Bag of Words (BoW) model differ from TF-IDF?**

**Answer:** The Bag of Words model represents text as a vector of word counts, focusing purely on word frequency without considering their importance. In contrast, TF-IDF considers both word frequency and how unique the word is in the corpus, assigning higher weights to less common words.

**3.** **3. What are the limitations of the Bag of Words (BoW) model?**

**Answer:**

* It ignores the order of words, losing the context and meaning.
* It creates sparse vectors, which can be inefficient for large vocabularies.
* It does not account for semantic relationships between words (e.g., synonyms).

#### **4.What is the significance of a Word Cloud in text analysis?**

**Answer:** A Word Cloud visually represents the most frequent or significant terms in a corpus, making it easy to identify key themes and topics at a glance. The size of each word corresponds to its frequency or importance.

**5.** **Which libraries are commonly used in Python for implementing TF-IDF, Bag of Words, and Word Cloud generation?**

**Answer:**

* For TF-IDF: TfidfVectorizer from the sklearn.feature\_extraction.text module.
* For Bag of Words: CountVectorizer from sklearn.
* For Word Cloud: The WordCloud library.

**Conslusion :** The practical implementation of text representation techniques like Bag of Words and TF-IDF, along with visualization using Word Clouds, is crucial for preprocessing and analyzing text data in Natural Language Processing. These methods help convert textual data into numerical formats, enabling machine learning models to extract meaningful insights and make predictions effectively. The Word Cloud further simplifies text exploration by visually summarizing prominent terms, aiding in better understanding of the corpus.