Assignment No: 7

Title of the Assignment:

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

Objective:

Students should be able to perform **Text Analysis** using TFIDF Algorithm

Prerequisite:

1. Basic of Python Programming
2. Basic of English language.

Contents for Theory:

1. Basic concepts of Text Analytics
2. Text Analysis Operations using natural language toolkit
3. Text Analysis Model using TF-IDF.
4. Bag of Words (BoW)

1. Basic concepts of Text Analytics

One of the most frequent types of day-to-day conversion is text communication. In our everyday routine, we chat, message, tweet, share status, email, create blogs, and offer opinions and criticism. All of these actions lead to a substantial amount of unstructured text being produced. It is critical to examine huge amounts of data in this sector of the online world and social media to determine people's opinions.

Text mining is also referred to as text analytics. Text mining is a process of exploring sizable textual data and finding patterns. Text Mining processes the text itself, while NLPprocesses with the underlying metadata. Finding frequency counts of words, length of the sentence, presence/absence of specific words is known as text mining. Natural language processing is one of the components of text mining. NLP helps identify sentiment, finding entities in the sentence, and category of blog/article. Text mining is preprocessed data for text analytics. In Text Analytics, statistical and machine learning algorithms are used to classify information.

2. Text Analysis Operations using natural language toolkit

NLTK(natural language toolkit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning and many more.

Analysing movie reviews is one of the classic examples to demonstrate a simple NLP Bag-of-words model, on movie reviews.

Tokenization:

Tokenization is the first step in text analytics. The process of breaking down a textparagraph into smaller chunks such as words or sentences is called Tokenization. Token is a single entity that is the building blocks for a sentence or paragraph.

- Sentence tokenization : split a paragraph into **list of sentences** using **sent_tokenize()** method
- Word tokenization : split a sentence into **list of words** using **word_tokenize()** method

Stop words removal

Stopwords considered as noise in the text. Text may contain stop words such as is,am, are, this, a, an, the, etc. In NLTK for removing stopwords, you need to create a list of stopwords and filter out your list of tokens from these words.

Stemming and Lemmatization

Stemming is a normalization technique where lists of tokenized words are converted into shortened root words to remove redundancy. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form.

A computer program that stems word may be called a stemmer.E.g.

A stemmer reduces the words like fishing, fished, and fisher to the stem fish.

The stem need not be a word, for example the Porter algorithm reduces, argue,argued, argues, arguing, and argus to the stem argu .

Lemmatization in NLTK is the algorithmic process of finding the lemma of a word depending on its meaning and context. Lemmatization usually refers to the morphological analysis of words, which aims to remove inflectional endings. It helps in returning the base or dictionary form of a word known as the lemma.

Eg. Lemma for studies is study

Lemmatization Vs Stemming

Stemming algorithm works by cutting the suffix from the word. In a broader sensecuts either the beginning or end of the word.

On the contrary, Lemmatization is a more powerful operation, and it takes into consideration morphological analysis of the words. It returns the lemma which is the base form of all its inflectional forms. In-depth linguistic knowledge is

required to create dictionaries and look for the proper form of the word. Stemming is a general operation while lemmatization is an intelligent operation where the proper form will be looked in the dictionary. Hence, lemmatization helps in forming better machine learning features.

2.2. POS Tagging

POS (Parts of Speech) tell us about grammatical information of words of the sentence by assigning specific token (Determiner, noun, adjective, adverb ,verb, Personal Pronoun etc.) as tag (DT,NN ,JJ,RB,VB,PRP etc) to each words.

Word can have more than one POS depending upon the context where it is used. We can use POS tags as statistical NLP tasks. It distinguishes a sense of word which is very helpful in text realization and infer semantic information from text for sentiment analysis.

Part-II

Text Analysis Model using TF-IDF.

Term frequency–inverse document frequency(TFIDF) , is a numerical statistic that is intended to reflect how important a word is to a document in a collection orcorpus.

- **Term Frequency (TF)**

It is a measure of the frequency of a word (w) in a document (d). TF is defined as the ratio of a

word's occurrence in a document to the total number of words in a document. The denominator term in the formula is to normalize since all the corpus documents are of different lengths.

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Example:

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

The initial step is to make a vocabulary of unique words and calculate TF for each document. TF will be more for words that frequently appear in a document and less for rare words in a document.

- **Inverse Document Frequency (IDF)**

It is the measure of the importance of a word. Term frequency (TF) does not consider the importance of words. Some words such as' of', 'and', etc. can be most frequently present but are of little significance. IDF provides weightage to each word based on its frequency in the corpus D.

$$IDF(w,D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

In our example, since we have two documents in the corpus, N=2.

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

Term Frequency — Inverse Document Frequency (TFIDF)

It is the product of TF and IDF.

TFIDF gives more weightage to the word that is rare in the corpus (all the documents).TFIDF provides more importance to the word that is more frequent in the document.

$$TFIDF\ (w, d, D) = TF(w, d) * IDF(w, D)$$

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

After applying TFIDF, text in A and B documents can be represented as a TFIDF vector of dimension equal to the vocabulary words. The value corresponding to each word representsthe importance of that word in a particular document.

TFIDF is the product of TF with IDF. Since TF values lie between 0 and 1, not using ***ln*** can result in high IDF for some words, thereby dominating the TFIDF. We don't want that, and therefore, we use ***ln*** so that the IDF should not completely dominate the TFIDF.

- **Disadvantage of TFIDF**

It is unable to capture the semantics. For example, funny and humorous are synonyms, but TFIDF does not capture that. Moreover, TFIDF can be computationally expensive if the vocabulary is vast.

3. Bag of Words (BoW)

Machine learning algorithms cannot work with raw text directly. Rather, the text must beconverted into vectors of numbers. In natural language processing, a common technique for extracting features from text is to place all of the words that occur in the text in abucket. This approach is called a bag of words model or BoW for short. It's referred toas a "bag" of words because any

information about the structure of the sentence is lost.

Algorithm for Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization:

Step 1: Download the required packages

```
nltk.download('punkt') nltk.download('stopwords')
nltk.download('wordnet') nltk.download('averaged_perceptron_tagger')
```

Step 2: Initialize the text

text= "Tokenization is the first step in text analytics. The process of breaking down a text paragraph into smaller chunksuch as words or sentences is called Tokenization."

Step 3: Perform Tokenization

```
#Sentence Tokenization
from nltk.tokenize import sent_tokenize
tokenized_text= sent_tokenize(text)print(tokenized_text)

#Word Tokenization
from nltk.tokenize import word_tokenizetokenized_word=word_tokenize(text)
print(tokenized_word)
```

Step 4: Removing Punctuations and Stop Word

```
# print stop words of English
from nltk.corpus import stopwords
stop_words=set(stopwords.words("english"))print(stop_words)

text= "How to remove stop words with NLTK library in Python?"text= re.sub('[^a-zA-Z]', ' ',text)
tokens = word_tokenize(text.lower())filtered_text=[]
for w in tokens:
if w not in stop_words: filtered_text.append(w)
print("Tokenized Sentence:",tokens) print("Filterd Sentence:",filtered_text)
```

Step 5 : Perform Stemming

```
from nltk.stem import PorterStemmer
e_words= ["wait", "waiting", "waited", "waits"]ps =PorterStemmer()
for w in e_words:
rootWord=ps.stem(w)print(rootWord)
```

Step 6: Perform Lemmatization

```
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer() text = "studies studying
cries cry" tokenization = nltk.word_tokenize(text)
```

```
for w in tokenization:
print("Lemma for {} is {}".format(w,
wordnet_lemmatizer.lemmatize(w)))
```

Step 7: Apply POS Tagging to text

```
import nltk
from nltk.tokenize import word_tokenize data="The pink sweater fit her
perfectly"words=word_tokenize(data)
for word in words: print(nltk.pos_tag([word]))
```

Algorithm for Create representation of document by calculating TFIDF

Step 1: Import the necessary libraries.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
```

Step 2: Initialize the Documents.

```
documentA = 'Jupiter is the largest Planet' documentB = 'Mars is the fourth planet from the Sun'
```

Step 3: Create BagofWords (BoW) for Document A and B.

```
bagOfWordsA = documentA.split(' ')bagOfWordsB = documentB.split(' ')
```

Step 4: Create Collection of Unique words from Document A and B.

```
uniqueWords = set(bagOfWordsA).union(set(bagOfWordsB))
```

Step 5: Create a dictionary of words and their occurrence for each document in the corpus

```
numOfWordsA = dict.fromkeys(uniqueWords, 0)
for word in bagOfWordsA: numOfWordsA[word] += 1
numOfWordsB = dict.fromkeys(uniqueWords, 0)for word in bagOfWordsB:
numOfWordsB[word] += 1
```

Step 6: Compute the term frequency for each of our documents.

```
def computeTF(wordDict, bagOfWords)tfDict = {}
bagOfWordsCount = len(bagOfWords) for word, count in
wordDict.items():
tfDict[word] = count / float(bagOfWordsCount)return tfDict
tfA = computeTF(numOfWordsA, bagOfWordsA)tfB =
computeTF(numOfWordsB, bagOfWordsB)
```

Step 7: Compute the term Inverse Document Frequency.

```
def computeIDF(documents):
import math
```

```
N = len(documents)

idfDict = dict.fromkeys(documents[0].keys(), 0)for document in documents:
for word, val in document.items()if val > 0:
idfDict[word] += 1

for word, val in idfDict.items(): idfDict[word] = math.log(N / float(val))
return idfDict
idfs = computeIDF([numOfWordsA, numOfWordsB])idfs
```

Step 8: Compute the term TF/IDF for all words.

```
def computeTFIDF(tfBagOfWords, idfs)tfidf = {}
for word, val in tfBagOfWords.items()tfidf[word] = val * idfs[word]
return tfidf
tfidfA = computeTFIDF(tfA, idfs)tfidfB = computeTFIDF(tfB,
idfs)
df = pd.DataFrame([tfidfA, tfidfB])df
```

Conclusion:

In this way we have done text data analysis using TF IDF algorithm.

Assignment Question:

1) Perform Stemming for text = "studies studying cries cry". Compare the results generated with Lemmatization. Comment on your answer how Stemming and Lemmatization differ from each other.

2) Write Python code for removing stop words from the below documents, conver the documents into lowercase and calculate the TF, IDF and TFIDF score for each document.

   documentA = 'Jupiter is the largest Planet' documentB = 'Mars is the fourth planet from the Sun'

.

In [29]:
```
import numpy as np
import nltk
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
```

In [30]:
```
text=['''Data science is an interdisciplinary academic field[1] that uses statistics, scienti

Data science also integrates domain knowledge from the underlying application domain (e.g., na

Data science is "a concept to unify statistics, data analysis, informatics, and their related

A data scientist is a professional who creates programming code and combines it with statistic
```

Tokenization

In [31]:
```
tokens=[word_tokenize(t.lower()) for t in text]

print(f'Tokens: {tokens}')
```

```
Tokens: [['data', 'science', 'is', 'an', 'interdisciplinary', 'academic', 'field', '[', '1',
']', 'that', 'uses', 'statistics', ',', 'scientific', 'computing', ',', 'scientific', 'method
s', ',', 'processing', ',', 'scientific', 'visualization', ',', 'algorithms', 'and', 'system
s', 'to', 'extract', 'or', 'extrapolate', 'knowledge', 'from', 'potentially', 'noisy', ',', 's
tructured', ',', 'or', 'unstructured', 'data', '.', '[', '2', ']', 'data', 'science', 'also',
'integrates', 'domain', 'knowledge', 'from', 'the', 'underlying', 'application', 'domain',
'(', 'e.g.', ',', 'natural', 'sciences', ',', 'information', 'technology', ',', 'and', 'medici
ne', ')', '.', '[', '3', ']', 'data', 'science', 'is', 'multifaceted', 'and', 'can', 'be', 'de
scribed', 'as', 'a', 'science', ',', 'a', 'research', 'paradigm', ',', 'a', 'research', 'metho
d', ',', 'a', 'discipline', ',', 'a', 'workflow', ',', 'and', 'a', 'profession', '.', '[',
'4', ']', 'data', 'science', 'is', '``', 'a', 'concept', 'to', 'unify', 'statistics', ',', 'da
ta', 'analysis', ',', 'informatics', ',', 'and', 'their', 'related', 'methods', "''", 'to', '`
`', 'understand', 'and', 'analyze', 'actual', 'phenomena', "''", 'with', 'data', '.', '[',
'5', ']', 'it', 'uses', 'techniques', 'and', 'theories', 'drawn', 'from', 'many', 'fields', 'w
ithin', 'the', 'context', 'of', 'mathematics', ',', 'statistics', ',', 'computer', 'science',
',', 'information', 'science', ',', 'and', 'domain', 'knowledge', '.', '[', '6', ']', 'howeve
r', ',', 'data', 'science', 'is', 'different', 'from', 'computer', 'science', 'and', 'informat
ion', 'science', '.', 'turing', 'award', 'winner', 'jim', 'gray', 'imagined', 'data', 'scienc
e', 'as', 'a', '``', 'fourth', 'paradigm', "''", 'of', 'science', '(', 'empirical', ',', 'theo
retical', ',', 'computational', ',', 'and', 'now', 'data-driven', ')', 'and', 'asserted', 'tha
t', '``', 'everything', 'about', 'science', 'is', 'changing', 'because', 'of', 'the', 'impac
t', 'of', 'information', 'technology', "''", 'and', 'the', 'data', 'deluge', '.', '[', '7',
']', '[', '8', ']', 'a', 'data', 'scientist', 'is', 'a', 'professional', 'who', 'creates', 'pr
ogramming', 'code', 'and', 'combines', 'it', 'with', 'statistical', 'knowledge', 'to', 'summar
ize', 'data', '.', '[', '9', ']']]
```

POS Tagging

In [32]:
```
pos_tags=[nltk.pos_tag(token) for token in tokens]

print(f'Pos-Tagging: {pos_tags}')
```

```
Pos-Tagging: [[('data', 'NNS'), ('science', 'NN'), ('is', 'VBZ'), ('an', 'DT'), ('interdiscipl
inary', 'JJ'), ('academic', 'JJ'), ('field', 'NN'), ('[', 'VBD'), ('1', 'CD'), (']', 'NN'),
('that', 'WDT'), ('uses', 'VBZ'), ('statistics', 'NNS'), (',', ','), ('scientific', 'JJ'), ('c
omputing', 'NN'), (',', ','), ('scientific', 'JJ'), ('methods', 'NNS'), (',', ','), ('processi
ng', 'NN'), (',', ','), ('scientific', 'JJ'), ('visualization', 'NN'), (',', ','), ('algorithm
s', 'NN'), ('and', 'CC'), ('systems', 'NNS'), ('to', 'TO'), ('extract', 'VB'), ('or', 'CC'),
('extrapolate', 'VB'), ('knowledge', 'NN'), ('from', 'IN'), ('potentially', 'RB'), ('noisy',
'JJ'), (',', ','), ('structured', 'JJ'), (',', ','), ('or', 'CC'), ('unstructured', 'JJ'), ('d
ata', 'NNS'), ('.', '.'), ('[', '$'), ('2', 'CD'), (']', 'NNP'), ('data', 'NNS'), ('science',
'NN'), ('also', 'RB'), ('integrates', 'VBZ'), ('domain', 'VBP'), ('knowledge', 'NN'), ('from',
'IN'), ('the', 'DT'), ('underlying', 'VBG'), ('application', 'NN'), ('domain', 'NN'), ('(',
'('), ('e.g.', 'NN'), (',', ','), ('natural', 'JJ'), ('sciences', 'NNS'), (',', ','), ('inform
ation', 'NN'), ('technology', 'NN'), (',', ','), ('and', 'CC'), ('medicine', 'NN'), (')',
')'), ('.', '.'), ('[', '$'), ('3', 'CD'), (']', 'NNP'), ('data', 'NNS'), ('science', 'NN'),
('is', 'VBZ'), ('multifaceted', 'VBN'), ('and', 'CC'), ('can', 'MD'), ('be', 'VB'), ('describe
d', 'VBN'), ('as', 'IN'), ('a', 'DT'), ('science', 'NN'), (',', ','), ('a', 'DT'), ('researc
h', 'NN'), ('paradigm', 'NN'), (',', ','), ('a', 'DT'), ('research', 'NN'), ('method', 'NN'),
(',', ','), ('a', 'DT'), ('discipline', 'NN'), (',', ','), ('a', 'DT'), ('workflow', 'NN'),
(',', ','), ('and', 'CC'), ('a', 'DT'), ('profession', 'NN'), ('.', '.'), ('[', 'CC'), ('4',
'CD'), (']', 'NN'), ('data', 'NNS'), ('science', 'NN'), ('is', 'VBZ'), ('``', '``'), ('a', 'D
T'), ('concept', 'NN'), ('to', 'TO'), ('unify', 'VB'), ('statistics', 'NNS'), (',', ','), ('da
ta', 'NNS'), ('analysis', 'NN'), (',', ','), ('informatics', 'NNS'), (',', ','), ('and', 'C
C'), ('their', 'PRP$'), ('related', 'JJ'), ('methods', 'NNS'), ("''", "''"), ('to', 'TO'), ('`
`', '``'), ('understand', 'VB'), ('and', 'CC'), ('analyze', 'VB'), ('actual', 'JJ'), ('phenome
na', 'NNS'), ("''", "''"), ('with', 'IN'), ('data', 'NNS'), ('.', '.'), ('[', '$'), ('5', 'C
D'), (']', 'NN'), ('it', 'PRP'), ('uses', 'VBZ'), ('techniques', 'NNS'), ('and', 'CC'), ('theo
ries', 'NNS'), ('drawn', 'VBP'), ('from', 'IN'), ('many', 'JJ'), ('fields', 'NNS'), ('within',
'IN'), ('the', 'DT'), ('context', 'NN'), ('of', 'IN'), ('mathematics', 'NNS'), (',', ','), ('s
tatistics', 'NNS'), (',', ','), ('computer', 'NN'), ('science', 'NN'), (',', ','), ('informati
on', 'NN'), ('science', 'NN'), (',', ','), ('and', 'CC'), ('domain', 'NN'), ('knowledge', 'N
N'), ('.', '.'), ('[', 'CC'), ('6', 'CD'), (']', 'NN'), ('however', 'RB'), (',', ','), ('dat
a', 'NNS'), ('science', 'NN'), ('is', 'VBZ'), ('different', 'JJ'), ('from', 'IN'), ('compute
r', 'NN'), ('science', 'NN'), ('and', 'CC'), ('information', 'NN'), ('science', 'NN'), ('.',
'.'), ('turing', 'VBG'), ('award', 'JJ'), ('winner', 'NN'), ('jim', 'NN'), ('gray', 'NN'), ('i
magined', 'VBD'), ('data', 'NNS'), ('science', 'NN'), ('as', 'IN'), ('a', 'DT'), ('``', '``'),
('fourth', 'JJ'), ('paradigm', 'NN'), ("''", "''"), ('of', 'IN'), ('science', 'NN'), ('(',
'('), ('empirical', 'JJ'), (',', ','), ('theoretical', 'JJ'), (',', ','), ('computational', 'J
J'), (',', ','), ('and', 'CC'), ('now', 'RB'), ('data-driven', 'JJ'), (')', ')'), ('and', 'C
C'), ('asserted', 'VBD'), ('that', 'IN'), ('``', '``'), ('everything', 'NN'), ('about', 'IN'),
('science', 'NN'), ('is', 'VBZ'), ('changing', 'VBG'), ('because', 'IN'), ('of', 'IN'), ('th
e', 'DT'), ('impact', 'NN'), ('of', 'IN'), ('information', 'NN'), ('technology', 'NN'), ("''",
"''"), ('and', 'CC'), ('the', 'DT'), ('data', 'NNS'), ('deluge', 'NN'), ('.', '.'), ('[', 'C
C'), ('7', 'CD'), (']', 'JJ'), ('[', '$'), ('8', 'CD'), (']', 'CC'), ('a', 'DT'), ('data', 'N
N'), ('scientist', 'NN'), ('is', 'VBZ'), ('a', 'DT'), ('professional', 'JJ'), ('who', 'WP'),
('creates', 'VBZ'), ('programming', 'VBG'), ('code', 'NN'), ('and', 'CC'), ('combines', 'NN
S'), ('it', 'PRP'), ('with', 'IN'), ('statistical', 'JJ'), ('knowledge', 'NN'), ('to', 'TO'),
('summarize', 'VB'), ('data', 'NNS'), ('.', '.'), ('[', '$'), ('9', 'CD'), (']', 'NN')]]
```

Stop Word Removal

In [33]:

```
import string

stop_words = set(stopwords.words("english"))

no_stop_word = [[word for word in token_list if word not in stop_words and word not in string
print("Tokens after stop words removal:", no_stop_word)
```

```
Tokens after stop words removal: [['data', 'science', 'interdisciplinary', 'academic', 'fiel
d', '1', 'uses', 'statistics', 'scientific', 'computing', 'scientific', 'methods', 'processin
g', 'scientific', 'visualization', 'algorithms', 'systems', 'extract', 'extrapolate', 'knowled
ge', 'potentially', 'noisy', 'structured', 'unstructured', 'data', '2', 'data', 'science', 'al
so', 'integrates', 'domain', 'knowledge', 'underlying', 'application', 'domain', 'e.g.', 'natu
ral', 'sciences', 'information', 'technology', 'medicine', '3', 'data', 'science', 'multifacet
ed', 'described', 'science', 'research', 'paradigm', 'research', 'method', 'discipline', 'work
flow', 'profession', '4', 'data', 'science', '```', 'concept', 'unify', 'statistics', 'data',
'analysis', 'informatics', 'related', 'methods', "'", '```', 'understand', 'analyze', 'actua
l', 'phenomena', "'", 'data', '5', 'uses', 'techniques', 'theories', 'drawn', 'many', 'field
s', 'within', 'context', 'mathematics', 'statistics', 'computer', 'science', 'information', 's
cience', 'domain', 'knowledge', '6', 'however', 'data', 'science', 'different', 'computer', 's
cience', 'information', 'science', 'turing', 'award', 'winner', 'jim', 'gray', 'imagined', 'da
ta', 'science', '```', 'fourth', 'paradigm', "'", 'science', 'empirical', 'theoretical', 'comp
utational', 'data-driven', 'asserted', '```', 'everything', 'science', 'changing', 'impact', 'i
nformation', 'technology', "'", 'data', 'deluge', '7', '8', 'data', 'scientist', 'professiona
l', 'creates', 'programming', 'code', 'combines', 'statistical', 'knowledge', 'summarize', 'da
ta', '9']]
```

Stemming

In [34]:
```
stemmer=PorterStemmer()
stemmed_tokens=[[stemmer.stem(words) for words in token] for token in tokens]
print("Stemming Token: ",stemmed_tokens)
```

```
Stemming Token:  [['data', 'scienc', 'is', 'an', 'interdisciplinari', 'academ', 'field', '[',
'1', ']', 'that', 'use', 'statist', ',', 'scientif', 'comput', ',', 'scientif', 'method', ',',
'process', ',', 'scientif', 'visual', ',', 'algorithm', 'and', 'system', 'to', 'extract', 'o
r', 'extrapol', 'knowledg', 'from', 'potenti', 'noisi', ',', 'structur', ',', 'or', 'unstructu
r', 'data', '.', '[', '2', ']', 'data', 'scienc', 'also', 'integr', 'domain', 'knowledg', 'fro
m', 'the', 'underli', 'applic', 'domain', '(', 'e.g.', ',', 'natur', 'scienc', ',', 'inform',
'technolog', ',', 'and', 'medicin', ')', '.', '[', '3', ']', 'data', 'scienc', 'is', 'multifac
et', 'and', 'can', 'be', 'describ', 'as', 'a', 'scienc', ',', 'a', 'research', 'paradigm',
',', 'a', 'research', 'method', ',', 'a', 'disciplin', ',', 'a', 'workflow', ',', 'and', 'a',
'profess', '.', '[', '4', ']', 'data', 'scienc', 'is', '```', 'a', 'concept', 'to', 'unifi', 's
tatist', ',', 'data', 'analysi', ',', 'informat', ',', 'and', 'their', 'relat', 'method',
"'", 'to', '```', 'understand', 'and', 'analyz', 'actual', 'phenomena', "'", 'with', 'data',
'.', '[', '5', ']', 'it', 'use', 'techniqu', 'and', 'theori', 'drawn', 'from', 'mani', 'fiel
d', 'within', 'the', 'context', 'of', 'mathemat', ',', 'statist', ',', 'comput', 'scienc',
',', 'inform', 'scienc', ',', 'and', 'domain', 'knowledg', '.', '[', '6', ']', 'howev', ',',
'data', 'scienc', 'is', 'differ', 'from', 'comput', 'scienc', 'and', 'inform', 'scienc', '.',
'ture', 'award', 'winner', 'jim', 'gray', 'imagin', 'data', 'scienc', 'as', 'a', '```', 'fourt
h', 'paradigm', "'", 'of', 'scienc', '(', 'empir', ',', 'theoret', ',', 'comput', ',', 'and',
'now', 'data-driven', ')', 'and', 'assert', 'that', '```', 'everyth', 'about', 'scienc', 'is',
'chang', 'becaus', 'of', 'the', 'impact', 'of', 'inform', 'technolog', "'", 'and', 'the', 'da
ta', 'delug', '.', '[', '7', ']', '[', '8', ']', 'a', 'data', 'scientist', 'is', 'a', 'profess
ion', 'who', 'creat', 'program', 'code', 'and', 'combin', 'it', 'with', 'statist', 'knowledg',
'to', 'summar', 'data', '.', '[', '9', ']']]
```

Lemmatization

In [35]:
```
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [[lemmatizer.lemmatize(word) for word in token] for token in tokens]
print("Lemmatized Tokens:", lemmatized_tokens)
```

```
Lemmatized Tokens: [['data', 'science', 'is', 'an', 'interdisciplinary', 'academic', 'field',
'[', '1', ']', 'that', 'us', 'statistic', ',', 'scientific', 'computing', ',', 'scientific',
'method', ',', 'processing', ',', 'scientific', 'visualization', ',', 'algorithm', 'and', 'sys
tem', 'to', 'extract', 'or', 'extrapolate', 'knowledge', 'from', 'potentially', 'noisy', ',',
'structured', ',', 'or', 'unstructured', 'data', '.', '[', '2', ']', 'data', 'science', 'als
o', 'integrates', 'domain', 'knowledge', 'from', 'the', 'underlying', 'application', 'domain',
'(', 'e.g.', ',', 'natural', 'science', ',', 'information', 'technology', ',', 'and', 'medicin
e', ')', '.', '[', '3', ']', 'data', 'science', 'is', 'multifaceted', 'and', 'can', 'be', 'des
cribed', 'a', 'a', 'science', ',', 'a', 'research', 'paradigm', ',', 'a', 'research', 'metho
d', ',', 'a', 'discipline', ',', 'a', 'workflow', ',', 'and', 'a', 'profession', '.', '[',
'4', ']', 'data', 'science', 'is', '``', 'a', 'concept', 'to', 'unify', 'statistic', ',', 'dat
a', 'analysis', ',', 'informatics', ',', 'and', 'their', 'related', 'method', "''", 'to', '`
`', 'understand', 'and', 'analyze', 'actual', 'phenomenon', "''", 'with', 'data', '.', '[',
'5', ']', 'it', 'us', 'technique', 'and', 'theory', 'drawn', 'from', 'many', 'field', 'withi
n', 'the', 'context', 'of', 'mathematics', ',', 'statistic', ',', 'computer', 'science', ',',
'information', 'science', ',', 'and', 'domain', 'knowledge', '.', '[', '6', ']', 'however',
',', 'data', 'science', 'is', 'different', 'from', 'computer', 'science', 'and', 'informatio
n', 'science', '.', 'turing', 'award', 'winner', 'jim', 'gray', 'imagined', 'data', 'science',
'a', 'a', '``', 'fourth', 'paradigm', "''", 'of', 'science', '(', 'empirical', ',', 'theoretic
al', ',', 'computational', ',', 'and', 'now', 'data-driven', ')', 'and', 'asserted', 'that',
'``', 'everything', 'about', 'science', 'is', 'changing', 'because', 'of', 'the', 'impact', 'o
f', 'information', 'technology', "''", 'and', 'the', 'data', 'deluge', '.', '[', '7', ']',
'[', '8', ']', 'a', 'data', 'scientist', 'is', 'a', 'professional', 'who', 'creates', 'program
ming', 'code', 'and', 'combine', 'it', 'with', 'statistical', 'knowledge', 'to', 'summarize',
'data', '.', '[', '9', ']']]
```

Term frequence and Inverse Document frequency

In [36]:
```python
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()
X=vectorizer.fit_transform(text)
vectorizer.get_feature_names_out()
```

Out[36]:
```
array(['about', 'academic', 'actual', 'algorithms', 'also', 'an',
       'analysis', 'analyze', 'and', 'application', 'as', 'asserted',
       'award', 'be', 'because', 'can', 'changing', 'code', 'combines',
       'computational', 'computer', 'computing', 'concept', 'context',
       'creates', 'data', 'deluge', 'described', 'different',
       'discipline', 'domain', 'drawn', 'driven', 'empirical',
       'everything', 'extract', 'extrapolate', 'field', 'fields',
       'fourth', 'from', 'gray', 'however', 'imagined', 'impact',
       'informatics', 'information', 'integrates', 'interdisciplinary',
       'is', 'it', 'jim', 'knowledge', 'many', 'mathematics', 'medicine',
       'method', 'methods', 'multifaceted', 'natural', 'noisy', 'now',
       'of', 'or', 'paradigm', 'phenomena', 'potentially', 'processing',
       'profession', 'professional', 'programming', 'related', 'research',
       'science', 'sciences', 'scientific', 'scientist', 'statistical',
       'statistics', 'structured', 'summarize', 'systems', 'techniques',
       'technology', 'that', 'the', 'their', 'theoretical', 'theories',
       'to', 'turing', 'underlying', 'understand', 'unify',
       'unstructured', 'uses', 'visualization', 'who', 'winner', 'with',
       'within', 'workflow'], dtype=object)
```

In [37]:
```python
tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform([" ".join(doc) for doc in lemmatized_tokens])
feature_names = tfidf_vectorizer.get_feature_names_out()
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=feature_names)

print("TF-IDF Representation:")
print(tfidf_df)
```

```
TF-IDF Representation:
      about  academic    actual  algorithm      also        an  analysis  \
0  0.035007  0.035007  0.035007   0.035007  0.035007  0.035007  0.035007

    analyze       and  application  ...  understand     unify  unstructured  \
0  0.035007  0.455091     0.035007  ...    0.035007  0.035007      0.035007

         us  visualization       who    winner      with    within  workflow
0  0.070014       0.035007  0.035007  0.035007  0.070014  0.035007  0.035007

[1 rows x 98 columns]
```