DSBDAL Assignment 07 - Text Analysis

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.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
     Mounted at /content/drive
```

```
import nltk

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

```
     [nltk_data] Downloading package punkt to /root/nltk_data...
     [nltk_data]   Unzipping tokenizers/punkt.zip.
     [nltk_data] Downloading package stopwords to /root/nltk_data...
     [nltk_data]   Unzipping corpora/stopwords.zip.
     [nltk_data] Downloading package averaged_perceptron_tagger to
     [nltk_data]     /root/nltk_data...
     [nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
     [nltk_data] Downloading package wordnet to /root/nltk_data...
     True
```

```
doc_file = open( "/content/drive/My Drive/DSBDL/Assignment7/doc_01" , "r" )
doc = doc_file.read()
doc_file.close()
print( doc )
```

```
     Between 2016 and 2019, the state forest department under the BJP government had launc
     In Pune Revenue Division, it was claimed the gram panchayats planted 1.7 crore saplir
     This year, the targets set by the forest department were comparatively modest. For ex
     In Pune Division – which comprises six talukas namely Maval, Mulshi, Daund, Indapur,
     The National Forest Policy aims and emphasizes at maintaining 33% of the country’s ge

     The plantation programme, which was announced in 2016 with the aim of planting 2 cror

     The 4 crore saplings for the year 2017 will be planted during the Vanmohotsav, July 1

     In a first of its kind, a 24-hour toll free helpline number 1926 called ‘Hello Forest

     In consonance of the public participation, the Maharashtra Forest Department has init

     An integrated drive has been set in place to ensure seamless and successful participa
```

```
import re

doc = re.sub('[\W_]+', ' ', doc )
```

```
doc
```

```
'Between 2016 and 2019 the state forest department under the BJP government had laun
ched Green Maharashtra drive with an aim to plant 50 crore trees across the state in
the four year period In October 2019 the government had claimed it had surpassed the
target by planting 33 crore trees in July September 2019 The Indian Express had foun
d that non forest agencies such as gram panchayats which were tasked with planting t
rees had not uploaded the mandatory audio visual proof of the tree plantation drives
on the specially created portal In Pune Revenue Division it was claimed the gram pan
chayats planted 1 7 crore saplings however no evidence was uploaded for 87 per cent
```

```
# word tokenization
word_tokens = nltk.word_tokenize( doc )
print( word_tokens )
```

```
['Between', '2016', 'and', '2019', 'the', 'state', 'forest', 'department', 'under', '
```

```
# sentence tokenization
sent_tokens = nltk.sent_tokenize( doc )
print( sent_tokens )
```

```
['Between 2016 and 2019 the state forest department under the BJP government had laun
```

```
stop_words = set(nltk.corpus.stopwords.words('english'))
word_tokens = [ token for token in word_tokens if token not in stop_words ]
print( word_tokens )
```

```
['Between', '2016', '2019', 'state', 'forest', 'department', 'BJP', 'government', 'la
```

```
tags = nltk.pos_tag( word_tokens )
print( tags )
```

```
[('Between', 'IN'), ('2016', 'CD'), ('2019', 'CD'), ('state', 'NN'), ('forest', 'JJS'
```

```
lemmatizer = nltk.stem.WordNetLemmatizer()
lemmatized_tokens = [ lemmatizer.lemmatize( token ) for token in word_tokens ]
print( lemmatized_tokens )
```

```
['Between', '2016', '2019', 'state', 'forest', 'department', 'BJP', 'government', 'la
```

```
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()
stemmed_tokens = [ stemmer.stem(token) for token in word_tokens ]
print( stemmed_tokens )
```

```
['between', '2016', '2019', 'state', 'forest', 'depart', 'bjp', 'govern', 'launch', '
```

```
with open( "/content/drive/My Drive/DSBDL/Assignment7/doc_02" , "r" ) as file:
    doc_2 = file.read()
doc_2 = re.sub('[\W_]+', ' ', doc_2 )
doc_2
```

```
'Millions of people in India took part in an annual tree planting drive Sunday More
than 250 million saplings were planted in a single day across the country s most pop
ulous state The campaign was led by Uttar Pradesh state government officials lawmake
rs and activists in a bid to reduce carbon emissions and combat climate change Where
were the trees planted The saplings were planted by volunteers in forests farms scho
ols and along riverbanks and highways We are committed to increasing the forest cove
r of Uttar Pradesh to over 15 of the total land area in the next five years said sta
te forest official Manoj Singh According to another government official the forest c
```

```
import numpy as np

def calc_term_freq(
    doc
):
    word_tokens = nltk.word_tokenize( doc )
    num_tokens = len( word_tokens )
    unique_tokens , freqs = np.unique( word_tokens , return_counts=True )
    tf = {}
    for token , freq in zip( unique_tokens , freqs ):
        tf[ token ] = freq / num_tokens
    return tf

tf = calc_term_freq( doc )
tf_2 = calc_term_freq( doc_2 )
```

```
import math

doc_1_tokens = nltk.word_tokenize( doc )
doc_2_tokens = nltk.word_tokenize( doc_2 )

def calc_idf():
    N = 2
    all_tokens = doc_1_tokens + doc_2_tokens
    idf = {}
    for token in all_tokens:
```

```
doc_1_repr = []
for token in doc_1_tokens:
    doc_1_repr.append( tf[ token ] * idf[token] )
doc_2_repr = []
for token in doc_2_tokens:
    doc_2_repr.append( tf_2[ token ] * idf[token] )
```

```
print( doc_1_repr )
```

```
[0.0, 0.0, -0.012003901226886446, -0.0026675336059747657, -0.02667533605974766, -0.00
```

```
print( doc_2_repr )
```

```
[0.0, -0.014898980350431239, 0.0, -0.011706341703910259, 0.0, 0.0, -0.001064212882173
```