ADVANCE AI LAB 01

ADITYA KAMLESH YADAV PRN : 20190802060

AIM : TO PERFORM SEGMENTATION AND FEATURE ENGINEERING ON TEXT DATA.

STEPS ::

1. DATA COLLECTION
2. SEGMENTATION
3. DATA PREPROCESSING
4. FEATURE ENGINEERING

FIND MY LAB REPORT ON GITHUB :: https://github.com/adityay186/ADVANCE-AI/lab_01.pdf

```
#using PyPDF2 library

import PyPDF2 as p

pdf1 = (p.PdfFileReader("/home/sample1.pdf")).getPage(0).extractText()
pdf2 = (p.PdfFileReader("/home/sample2.pdf")).getPage(0).extractText()
pdf3 = (p.PdfFileReader("/home/sample3.pdf")).getPage(0).extractText()

#splitting strings to list

pdf1 = pdf1.splitlines()
pdf2 = pdf2.splitlines()
pdf3 = pdf3.splitlines()

print(pdf1)
print(pdf2)
print(pdf3)
```

```
     ['Instructions for Adding Y our Logo & Address to AAO-HNSF Patient Handouts', 'CO-BRAN
     ['PDF Form Example', 'This is an example of a user fillable PDF form. Normally PDF is
     ['   (c) eaDocX Ltd 2012   ', '1  ', ' ', ' ', ' ', ' ', 'Get Lost!  ', 'A Document wh
```

```
#extracting title and merging them into single string
pdf1="".join(pdf1[4:7])
pdf2="".join(pdf2[3:5])
pdf3="".join(pdf3[4:6])

print(pdf1)
print(pdf2)
print(pdf3)


#removing stop words and pronounciation
import spacy
nlp = spacy.load("en_core_web_sm")

def process_data(txt):
  doc = nlp(txt)
  filtered_words = []
  for word in doc:
```

```
        continue
    filtered_words.append(word.lemma_)
  return " ".join(filtered_words)

data1 =process_data(pdf1)
data2 =process_data(pdf2)
data3 =process_data(pdf3)
#adding space and printing data
data1=data1[:25] + " " + data1[25:]
data1=data1[:60] + " " + data1[60:]
data3=data3[:27] + " " + data3[27:]
print(data1)
print(data2)
print(data3)


#feature extraction using TF-IDF Method

from sklearn.feature_extraction.text import TfidfVectorizer
#creating transformer
vectorizer = TfidfVectorizer()
#Train the Model by fitting the text_documents.
vectorizer.fit([data1,data2,data3])
print(vectorizer.vocabulary_)


vector1=vectorizer.transform([data1])
vector2=vectorizer.transform([data2])
vector3=vectorizer.transform([data3])
#encoded vector
print(vector1.toarray())
print(vector2.toarray())
print(vector3.toarray())
```

```
    [[1.]]
    [[0.]]
    [[0.]]
```

[[0.37380112 0. 0.37380112 0.37380112 0.28428538 0. 0.37380112 0. 0. 0. 0. 0.

0. 0. 0. 0.37380112 0.28428538 0.37380112

1. 0.]] [[0. 0. 0. 0. 0.37302199 0.

2. 0. 0.49047908 0.49047908 0. 0.

3. 0. 0. 0. 0.37302199 0.

4. 49047908 0.]] [[0. 0.33333333 0. 0. 0. 0.33333333

5. 0.33333333 0. 0. 0.33333333 0.33333333

The TF-IDF method gives us the importance of words present in the document.

CONCLUSION : Successfully extracted the title and journal name from the given PDF's and performed data preprocessing and feature engineering.