ADVANCE AI LAB ASSIGNMENT - 01

Name:Suman Kumar

PRN NO:20190802080

Aim: To perform segmentation and Feature Engineering on Text data.

Segment to Extract: Title and Journal Name.

Steps to perform:

1. Data Collection.
2. Segments Extraction GroupWise. You can consider synonyms if you don't find the relative keywords.
3. Data Preprocessing.
4. Feature Engineering

GIthub link : https://github.com/Mandeep3007/ADVANCE-AI-/20190802080_LAB_01.pdf

```
#installing required libraries
!pip install PyPDF2
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/p
Collecting PyPDF2
  Downloading PyPDF2-2.11.0-py3-none-any.whl (220 kB)
     |████████████████████████████████| 220 kB 2.1 MB/s
Requirement already satisfied: typing-extensions>=3.10.0.0 in /usr/local/lib/python3
```

To undo cell deletion use Ctrl+M Z or the Undo option in the Edit menu

```
#importing libraries
import PyPDF2 as p
```

```
#importing libraries
import PyPDF2 as p
#reading the pdfs and extracting the first page as title is required
pdf1 = (p.PdfFileReader("/content/file-example_01.pdf")).getPage(0).extractText()
pdf2 = (p.PdfFileReader("/content/Blockchain Techonology and Application -Surprise test Qu
pdf3 = (p.PdfFileReader("/content/Surprise Test_Suman_kumar_20190802080.pdf")).getPage(0).
```

Segment Extraction (Title and Journal Name)

```
#splitting strings to list
pdf1 = pdf1.splitlines()
pdf2 = pdf2.splitlines()
pdf3 = pdf3.splitlines()
```

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

```
 And more text. And more text. And more text. And more text. And more  text. And more
Subject Name:Blockchain Technology and Applications.
* SEC T ION-O1. MCe) A Disintmealia io.
```

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

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

```
 mo
je
C
```

Data Preprocessing & Feature Engineering

```
#Converting text to lowe case for feature extraction
```

To undo cell deletion use Ctrl+M Z or the Undo option in the Edit menu

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

```
 mo
je
c
```

Removing Stop words and punctuations

```
import spacy
nlp = spacy.load("en_core_web_sm")
```

```
def process_data(txt):
    doc = nlp(txt)
    filtered_words = []
    for word in doc:
      if word.is_stop or word.is_punct:
        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)
```

```
  mo
je
c
```

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.
```

To undo cell deletion use Ctrl+M Z or the Undo option in the Edit menu

```
{'mo': 1, 'je': 0}
```

Observation:

Words that appear frequently have the lowest values and unique words have the higest value

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

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

Observation

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

To undo cell deletion use Ctrl+M Z or the Undo option in the Edit menu