

## Exp 3 Manas Parab-40 CSE(DS) DLOC-NLP

## ▼ Library required

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex<=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
```

## ▼ Text

```
text = 'NGC 7319 is a highly active, blue-shifted emission-line galaxy and Seyfert Type 2 object situated in the Pegasus constellation, not far
```

```
text
```

```
↳ 'NGC 7319 is a highly active, blue-shifted emission-line galaxy and Seyfert Type 2 object situated in the Pegasus constellation, not far from the conjunction of Equuleus and Lacerta. Its estimated co-moving distance is about 280 million light-years away from our plane of observation.'
```

## ▼ Stopwords

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
True
```

```
stop_words = stopwords.words('english')
```

```
from nltk.tokenize import word_tokenize
nltk.download('punkt')
words = word_tokenize(text)

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
```

## ▼ Applying stop words

```
holder = list()
for w in words:
    if w not in set(stop_words):
        holder.append(w)
```

```
holder
```

```
['NGC',
 '7319',
 'highly',
 'active',
 ',',
 'blue-shifted',
 'emission-line',
 'galaxy',
 'Seyfert',
 'Type',
 '2',
 'object',
 'situated',
 'Pegasus',
 'constellation',
 ',',
 'far',
```

```
'conjunction',
'Equuleus',
'Lacerta',
',',
'It's',
'estimated',
'co-moving',
'distance',
'280',
'million',
'light-years',
'away',
'planet',
'.']
```

### ▼ List Comprehension for stop words

```
holder = [w for w in words if w not in set(stop_words)]
print(holder)
```

```
['NGC', '7319', 'highly', 'active', ',', 'blue-shifted', 'emission-line', 'galaxy', 'Seyfert', 'Type', '2', 'object', 'situated', 'Pegas
```

### ▼ Stemming

```
from nltk.stem import PorterStemmer, SnowballStemmer, LancasterStemmer
```

```
porter = PorterStemmer()
snow = SnowballStemmer(language = 'english')
lancastr = LancasterStemmer()
```

```
words = ['play', 'plays', 'played', 'playing', 'player']
```

### ▼ Porter Stemmer

```
porter_stemmed = list()
for w in words:
    stemmed_words = porter.stem(w)
    porter_stemmed.append(stemmed_words)
```

```
porter_stemmed
```

```
['play', 'play', 'play', 'play', 'player']
```

### ▼ Porter Stemmer List Comprehension

```
porter_stemmed = [porter.stem(x) for x in words]
print (porter_stemmed)
```

```
['play', 'play', 'play', 'play', 'player']
```

### ▼ Snowball Stemmer

```
snow_stemmed = list()
for w in words:
    stemmed_words = snow.stem(w)
    snow_stemmed.append(stemmed_words)
```

```
snow_stemmed
```

```
['play', 'play', 'play', 'play', 'player']
```

### ▼ Snowball Stemmer List Comprehension

23/08/2023, 14:37

```
snow_stemmed = [snow.stem(x) for x in words]
print (snow_stemmed)

['play', 'play', 'play', 'play', 'player']
```

#### ▼ Lancaster Stemmer

```
lancaster_stemmed = list()
for w in words:
    stemmed_words = lancaster.stem(w)
    lancaster_stemmed.append(stemmed_words)

lancaster_stemmed

['play', 'play', 'play', 'play', 'play']
```

#### ▼ Lancaster Stemmer List Comprehension

```
lancaster_stemmed = [lancaster.stem(x) for x in words]
print (lancaster_stemmed)

['play', 'play', 'play', 'play', 'play']
```

#### ▼ Lemmatization : This has a more expansive vocabulary than Stemming

```
from nltk.stem import WordNetLemmatizer
wordnet = WordNetLemmatizer()

nltk.download('wordnet')
lemmatized = [wordnet.lemmatize(x) for x in words]

[nltk_data] Downloading package wordnet to /root/nltk_data...

lemmatized

['play', 'play', 'played', 'playing', 'player']
```

0s completed at 14:31