Experiment 02 NLP DLOC Yash Sankhe 53 Dept. of CSE & DS VCET

Library required for Preprocessing

```
! pip install nltk
```

```
Requirement already satisfied: nltk in /usr/10ca1/1ib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/10ca1/1ib/python3.10/dist-packages (from nltk) (8.1.6)
Requirement already satisfied: joblib in /usr/10ca1/1ib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/10ca1/1ib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/10ca1/1ib/python3.10/dist-packages (from nltk) (4.66.1)
```

```
import nltk

nltk. download( ' punkt' )
```

```
[nltk_data] Downloading package punkt to / root/nltk_data. . .
[nltk_data] Package punkt is already up-to-date! True
```

Sentence Tokenization

```
from nltk.tokenize import sent_tokenize

text = 'Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing other stars 1 Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1,940 2,169 solar radii) . ' ,

text
```

```
„Stephenson 2-18 i s now known as being one of the largest, if not the current largest s tar ever discovered, surpassing other stars like VY Canis Majoris and IJY Scuti. \n q•t-ønhøncnn O - IR hac a radius a-F 1 ;0 solar rad •i * hAino 1 Than *Imncz-t- •t-høentir
```

```
sentences = sent_tokenize (text)

sentences
```

```
[ 'Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing other stars like VY Canis Majoris and IJY Scuti.' ,
'Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1,940 2,169 solar radii) .' ]
```

Word Tokenization

```
from nltk.tokenize import word_tokenize

words = word_tokenize (text)

words
```

```
[ 'Stephenson' ,
' 2-18 ' , „is
' „now' ,
' known' , „as'
'being' , „one '
' of' , 'the' ,
' largest' ,
' if ' ,
' not' ,
'the' , „current' ,
' largest' ,
' star' ,
```

```
 'ever' , ' discovered' ,
 'surpassing' , 'other' , 'stars ' ' like ' ,
                                                  'Canis' ,
 'Majoris ' ,
                                                   'and' ,
                                                   'UY' ,
                                                  'Scuti' ,
                                                                                  'Stephenson ' ,
 ' 2-18 ' ,
 'has' , 'a'
 ' radius'
 ,
 ' 2, 150
 ' , 'solar'
 , ' radii '
 ,
 'being' , ' larger' ,
                                                   'than' ,
                                                  ' almost' ,
                                                   'the' ,
                                                  ' entire '
 'orbit' , 'of' ,
                                                  'Saturn '
                                                 ' 1, 940 ' ,
                                                 ' 2, 169 ' ,
                                                  'solar' ,
                                                  ' radii '
```

```
for w in words:
    print (w)
```

2-18 is now known as being one Of the largest

if not the current largest star ever discovered

surpassing other stars like

Canis Majoris and IJY
Scuti

Stephenson 2-18 has a radius of 2, 150 solar
being

than
almost the entire orbit of Saturn

1, 940

2, 169 solar radii

Levels of Sentences Tokenization using Comprehension

sent_tokenize (text)

['Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing other stars like VY Canis Majoris and UY Scuti. ' ,
'Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1, 940 2,169 solar radii) .']

[word_tokenize (text) for t in sent_tokenize(text)]

[['Stephenson' ,
' 2-18 ' , '
is ' ,now' ,
' known '
'being' , ,one '
'Of' ,
'the' , ,largest' ,
' if ' ,
' not' ,
'the' , ,current' , ' largest' , ,star' ,
'ever' ,
' discovered' ,
,surpassing' ,
' other' ,
' stars' ,
' like ' ,

'Canis'
'Majoris ' , 'and' ,
'Scuti' ,

,Stephenson' ,
' 2-18 '
,
'has' ,
,a'
' radius '
' of' ,
' 2, 150
' , '
solar' ,
, radii '
,

,being' ,
' larger' ,

```
 'than' ,
 ' almost' ,
 'the' ,
 ' entire '
 'orbit' ,
 'Of' ,
 'Saturn '
 ' 1, 940
 ' ,
 ' 2, 169
 ' , '
 solar' ,
 ' radii '
 ,
```

```
from nltk.tokenize import wordpunct_tokenize

wordpunct_tokenize (text)
```

```
[ 'Stephenson' ,
 ' 2 ' ,
 ' 18 ' , ' is ' , 'now' ,
 ' known' , 'as'
 'being' ,
 'one' ,
 'the' , 'largest' ,
 ' if ' ,
 ' not' , 'the' , 'current' , 'largest' , ' star' , 'ever '
 ' discovered' ,
 'surpassing' ,
 ' other' ,
 ' stars '
 ' like ' ,
 'Canis'
 'Majoris ' ,
 'and' ,
 'UY' ,
 'Scuti' ,
 'Stephenson ' ,
 ' 2 ' ,
 ' 18 ' ,
 'has' , 'a ' , 'radius '
 'Of' ,
 ' 150 ' ,
 'solar' , ' radii ' ,
 'being' ,
 ' larger' ,
 'than' ,
 ' almost' ,
 'the' , 'entire' , 'orbit' , 'Of' ,
 'Saturn '
```

Filteration of Text by converting into lower case

```
text . lower()
```

```
' stephenson 2-18 i s now known as being one of the largest, if not the current lar gest
star ever discovered, surpassing other stars like vy canis majoris and uy sc c tanhøncnn
                                7 -IR hac: a radius of 2 1 SR cm I n r'      no I          thn
```

```
text . upper ( )
```

```
'STEPHENSON 2-18 IS NOW KNOWN AS BEING ONE OF THE LARGEST, IF NOT THE
CURRENT LAR GEST STAR EVER DISCOVERED, SURPASSING OTHER STARS LIKE VY
CANIS MAJORIS AND LJY SC
```

```
I ITT \ n                STEPHENSON '-IR          ARADTUS OF 2 ISA qnl AR RAnTT RFTNG I          THA
```

Colab paid products - Cancel contracts here s/ Os completed at 14:33