NLP Experiment 02 ~ 39_Sanskruti Nijai

Library required for Preprocessing

```
!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) (8.1.6)
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)
```

```
import nltk
```

```
nltk.download()
```

```
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader> d

Download which package (l=list; x=cancel)?
  Identifier> punkt
    Downloading package punkt to /root/nltk_data...
      Unzipping tokenizers/punkt.zip.

---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-3-6e230a00a763> in <cell line: 1>()
----> 1 nltk.download()

                                 4 frames
/usr/local/lib/python3.10/dist-packages/ipykernel/kernelbase.py in _input_request(self, prompt, ident, parent, password)
    893             except KeyboardInterrupt:
    894                 # re-raise KeyboardInterrupt, to truncate traceback
--> 895                 raise KeyboardInterrupt("Interrupted by user") from None
    896             except Exception as e:
    897                 self.log.warning("Invalid Message:", exc_info=True)

KeyboardInterrupt: Interrupted by user
```

SEARCH STACK OVERFLOW

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 sta
        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 is now known as being one of the largest, if not the current larges
t star ever discovered, surpassing other stars like VY Canis Majoris and UY Scuti.\n
Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the enti
```

```
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 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 Tokenization

```
from nltk.tokenize import word_tokenize


words = word_tokenize (text)


words
```

```
      '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',
      ')',
      '.']
```

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

```
     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
`
```

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',
  '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',
 '\)'
```

```
from nltk.tokenize import wordpunct_tokenize
```

```
wordpunct_tokenize (text)
```

```
'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',
').']
```

Filteration of Text by converting into lower case

```
text.lower()
```

```
'stephenson 2-18 is now known as being one of the largest, if not the current larges
t star ever discovered, surpassing other stars like vy canis majoris and uy scuti.\n
stephenson 2-18 has a radius of 2.150 solar radii, being larger than almost the enti
```

```
text.upper()
```

```
'STEPHENSON 2-18 IS NOW KNOWN AS BEING ONE OF THE LARGEST, IF NOT THE CURRENT LARGES
T STAR EVER DISCOVERED, SURPASSING OTHER STARS LIKE VY CANIS MAJORIS AND UY SCUTI.\n
STEPHENSON 2-18 HAS A RADIUS OF 2.150 SOLAR RADII. BEING LARGER THAN ALMOST THE ENTI
```