Experiment No.2
Apply Tokenization on given English and Indian Language Text
Date of Performance:
Date of Submission:

Aim: Apply Tokenization on given English and Indian Language Text

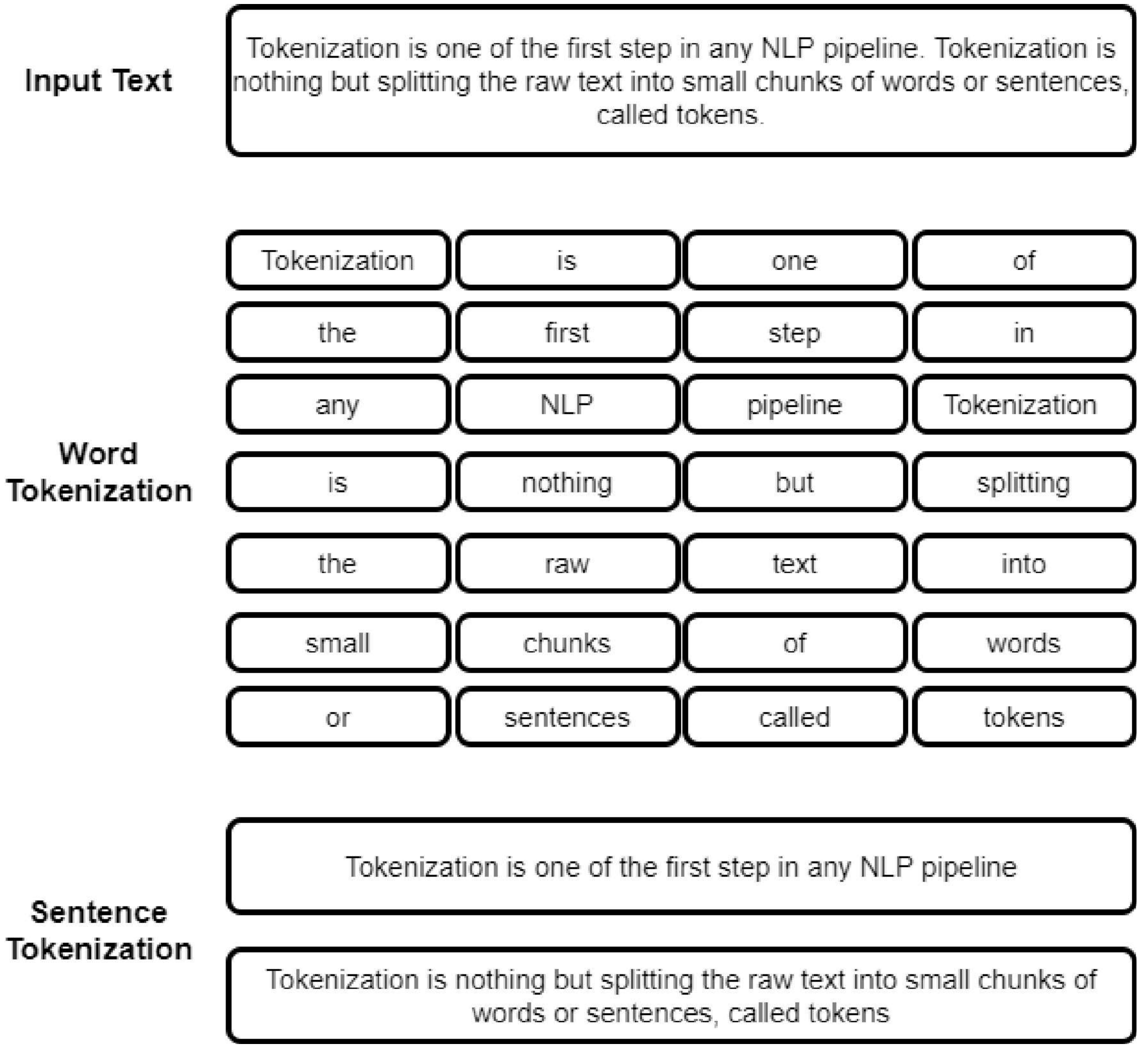
Objective: Able to perform sentence and word tokenization for the given input text for English and Indian Langauge.

Theory:

Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens. If the text is split into words, then its called as 'Word Tokenization' and if it's split into sentences then its called as 'Sentence Tokenization'. Generally 'space' is used to perform the word tokenization and characters like 'periods, exclamation point and newline char are used for Sentence Tokenization. We have to choose the appropriate method as per the task in hand. While performing the tokenization few characters like spaces, punctuations are ignored and will not be the part of final list of tokens.

Why Tokenization is Required?

Every sentence gets its meaning by the words present in it. So by analyzing the words present in the text we can easily interpret the meaning of the text. Once we have a list of words we can also use statistical tools and methods to get more insights into the text. For example, we can use word count and word frequency to find out important of word in that sentence or document.



Input Text

Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens.

Word Tokenization

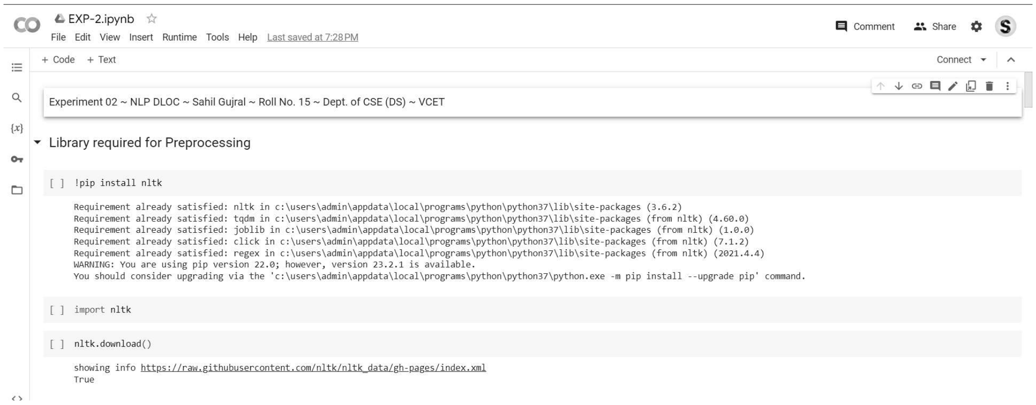
Tokenization	is	one	of
the	first	step	in
any	NLP	pipeline	Tokenization
is	nothing	but	splitting
the	raw	text	into
small	chunks	of	words
or	sentences	called	tokens

Sentence Tokenization

Tokenization is one of the first step in any NLP pipeline

Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens

Code:


EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment Share
+ Code + Text
Connect

Experiment 02 ~ NLP DLOC ~ Sahil Gujral ~ Roll No. 15 ~ Dept. of CSE (DS) ~ VCET

Library required for Preprocessing

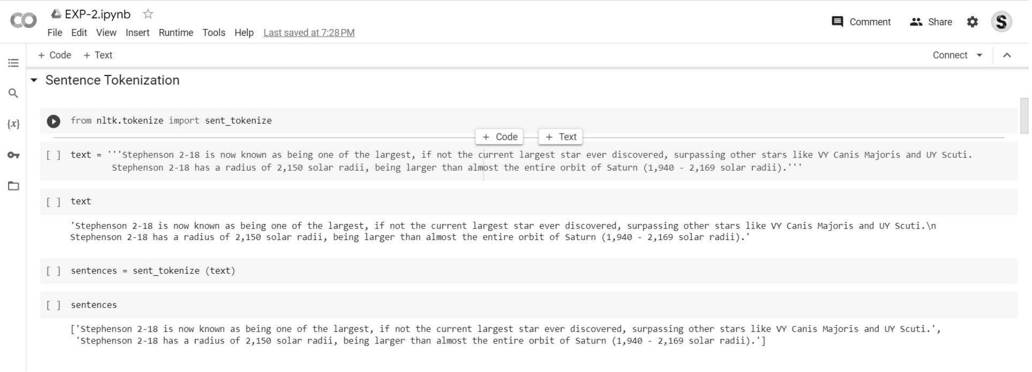
```
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\admin\appdata\local\programs\python\python37\lib\site-packages (3.6.2)
Requirement already satisfied: tqdm in c:\users\admin\appdata\local\programs\python\python37\lib\site-packages (from nltk) (4.60.0)
Requirement already satisfied: joblib in c:\users\admin\appdata\local\programs\python\python37\lib\site-packages (from nltk) (1.0.0)
Requirement already satisfied: click in c:\users\admin\appdata\local\programs\python\python37\lib\site-packages (from nltk) (7.1.2)
Requirement already satisfied: regex in c:\users\admin\appdata\local\programs\python\python37\lib\site-packages (from nltk) (2021.4.4)
WARNING: You are using pip version 22.0; however, version 23.2.1 is available.
You should consider upgrading via the 'c:\users\admin\appdata\local\programs\python\python37\python.exe -m pip install --upgrade pip' command.
```

```
import nltk
```

```
nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
True
```


EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment Share
+ Code + Text
Connect

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 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).'''
```

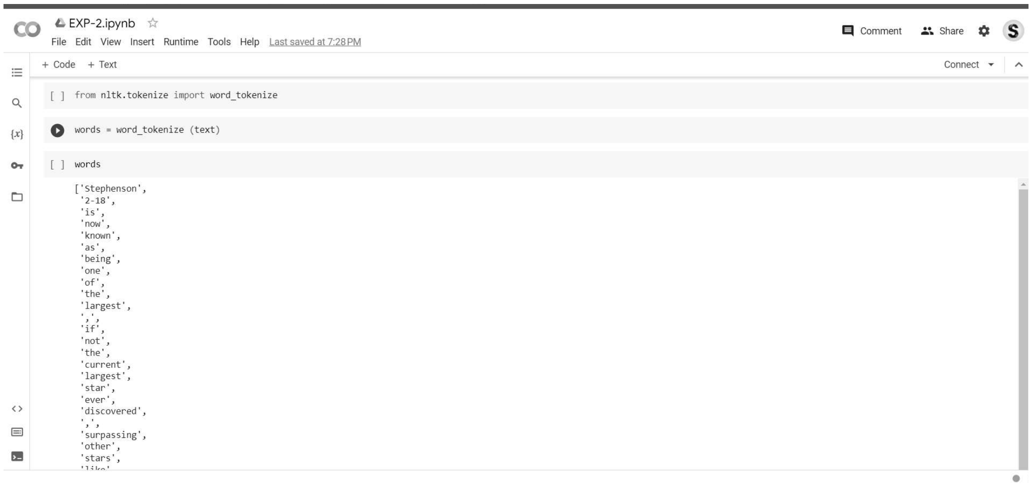
```
text
```

```
'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.\n
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).'
```

```
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).']
```

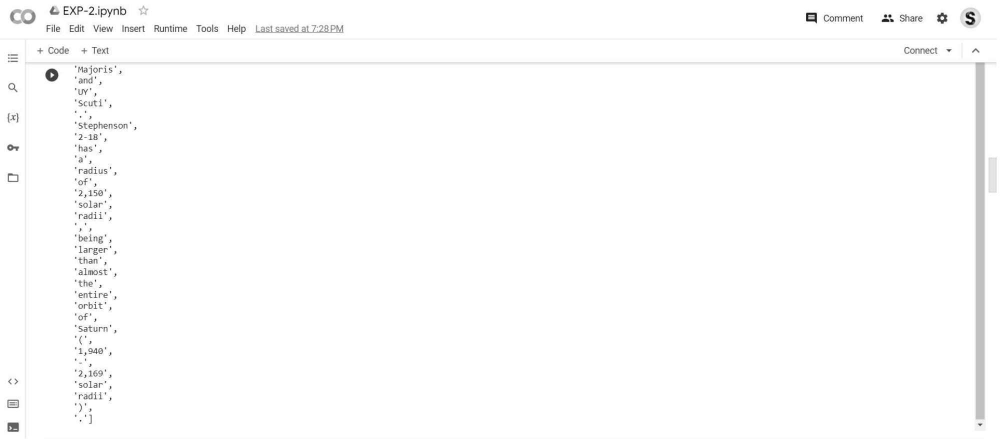

EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment Share
+ Code + Text
Connect

```
from nltk.tokenize import word_tokenize
```

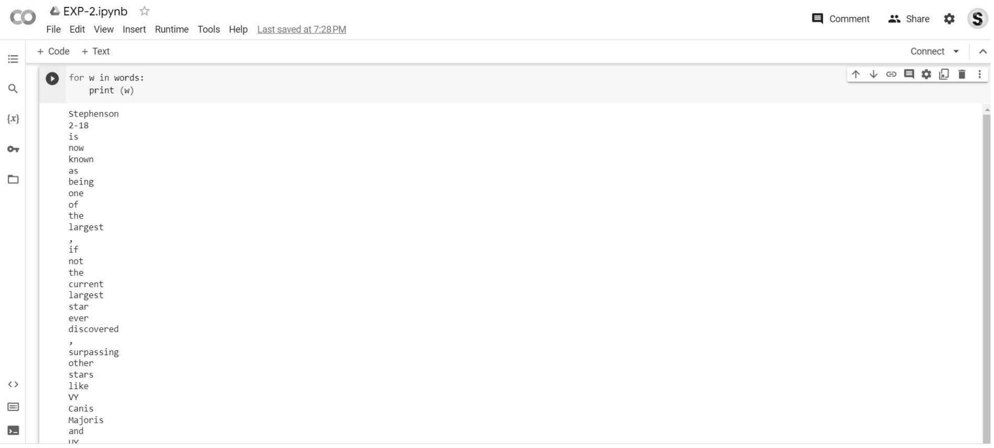
```
words = word_tokenize (text)
```

```
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',
```

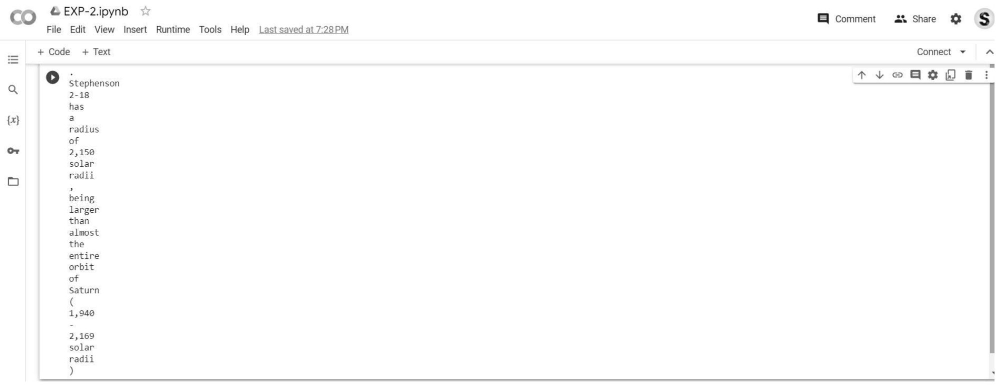

EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment
Share
+ Code + Text
Connect

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

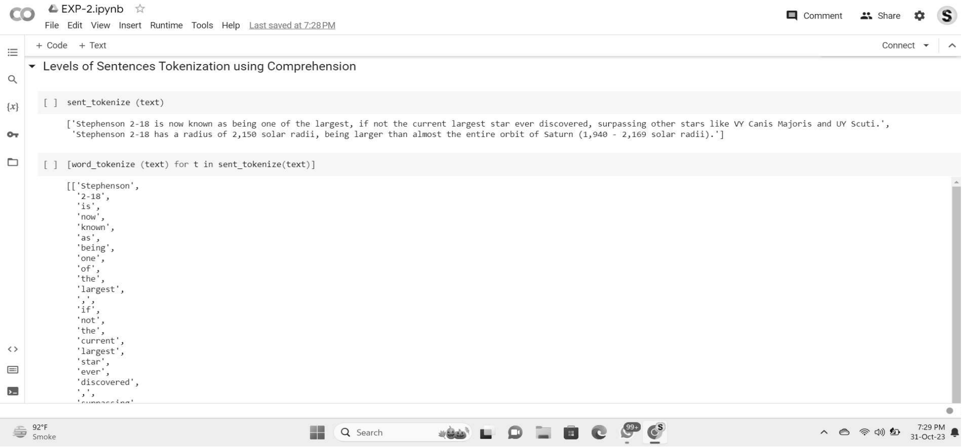

EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment
Share
+ Code + Text
Connect

```
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
```

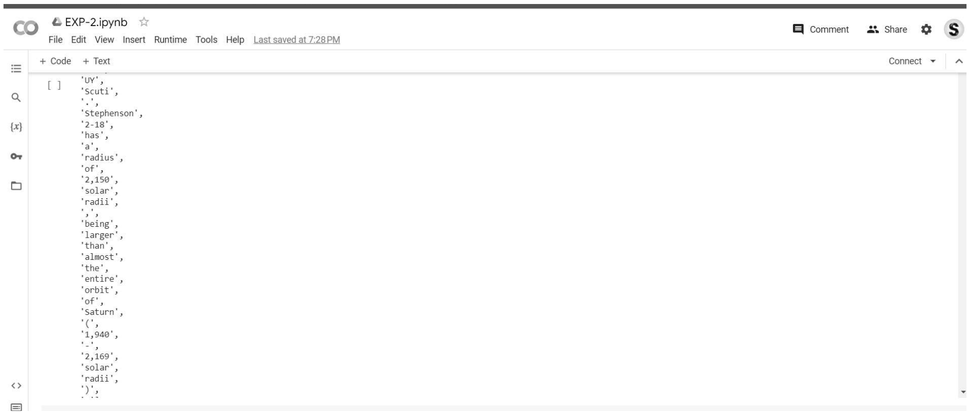

EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment
Share
+ Code + Text
Connect

```
.
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
)
```

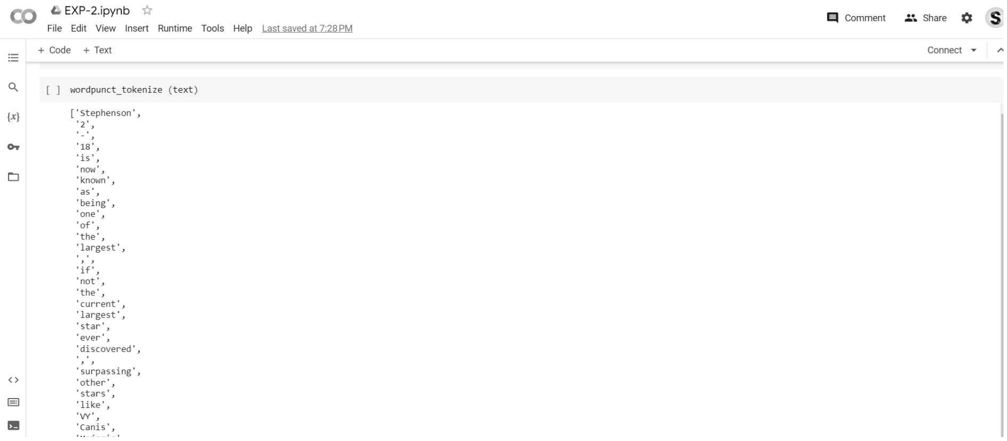


Levels of Sentences Tokenization using Comprehension

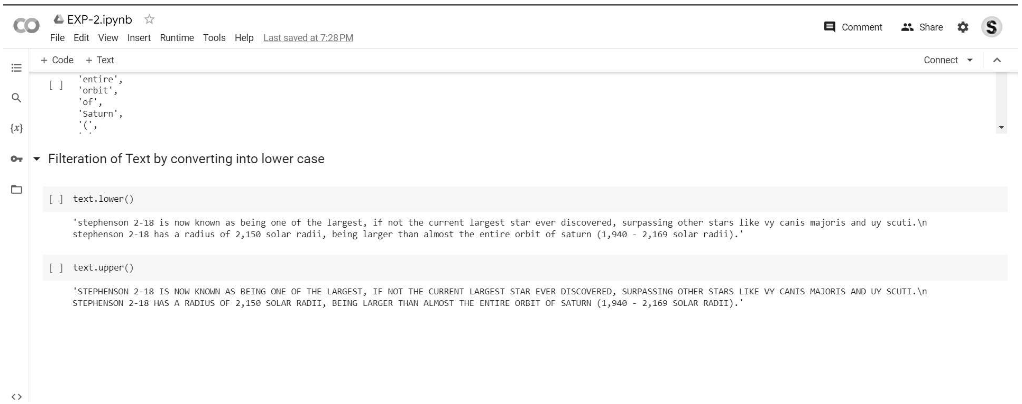
```
[ ] sent_tokenize (text)

    ['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'
```


EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment
Share
+ Code + Text
Connect

```
[ ]   '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',
      ')',
```


EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment Share
+ Code + Text
Connect

```
[ ] wordpunct_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',
```


EXP-2.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:28 PM
Comment Share
+ Code + Text
Connect

```
     'entire',
     'orbit',
     'of',
     'Saturn',
     '(',
```

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 largest 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 entire orbit of saturn (1,940 - 2,169 solar radii).'

[ ] text.upper()

    '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.\n
    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).'
```

Conclusion:

Comment on the tools used for tokenization of Indian language input.