Tokenisation in Hindi

[1]:
```
import nltk
```

[3]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Imart\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

[3]:
```
True
```

[8]:
```
corpus = ''' मेरा नाम अवनीश है और मुझे कुछ एमएल अवधारणाएँ सिखाना पसंद है। और मुझे इसमें मज़ा आता है। '''
```

[5]:
```
print('Corpus:',corpus)
```

```
Corpus:  मेरा नाम अवनीश है और मुझे कुछ एमएल अवधारणाएँ सिखाना पसंद है।
और मुझे इसमें मज़ा आता है।
```

[6]:
```
nltk.word_tokenize(corpus) # corpus -> words
```

[6]:
```
['मेरा',
 'नाम',
 'अवनीश',
 'है',
 'और',
 'मुझे',
 'कुछ',
 'एमएल',
 'अवधारणाएँ',
 'सिखाना',
 'पसंद',
 'है।',
 'और',
 'मुझे',
 'इसमें',
 'मज़ा',
 'आता',
 'है।']
```

[10]:
```
nltk.sent_tokenize(corpus) #corpus cannot be converted to setence.
```

[10]:
```
[' मेरा नाम अवनीश है और मुझे कुछ एमएल अवधारणाएँ सिखाना पसंद है। और मुझे इसमें मज़ा आता है।']
```

using the indic-nlp-library.

[13]:
```
from indicnlp.tokenize import indic_tokenize
words = indic_tokenize.trivial_tokenize(corpus,lang='hi')
```

[14]:
```
print('Tokenised words:',words)
```

```
Tokenised words: ['मेरा', 'नाम', 'अवनीश', 'है', 'और', 'मुझे', 'कुछ', 'एमएल', 'अवधारणाएँ', 'सिखाना', 'पसंद', 'है', '।', 'और', 'मुझे', 'इसमें', 'मज़ा', 'आता', 'है', '।']
```

[15]:
```
from indicnlp.tokenize.sentence_tokenize import sentence_split
sentences = sentence_split(corpus,lang='hi')
```

[17]:
```
print('sentences  or documents :',sentences)
```

```
sentences  or documents : ['मेरा नाम अवनीश है और मुझे कुछ एमएल अवधारणाएँ सिखाना पसंद है।', 'और मुझे इसमें मज़ा आता है।']
```

[19]:
```
lower_words = [word.lower() for word in words]
print('lowers words :',lower_words)
```

```
lowers words : ['मेरा', 'नाम', 'अवनीश', 'है', 'और', 'मुझे', 'कुछ', 'एमएल', 'अवधारणाएँ', 'सिखाना', 'पसंद', 'है', '।', 'और', 'मुझे', 'इसमें', 'मज़ा', 'आता', 'है', '।']
```

[20]:
```
vocabs = set(lower_words)
print('No of words:',len(lower_words))
print('vocabs :',vocabs)
print('No of Unique Words or Vocabs:',len(vocabs))
```

```
No of words: 20
vocabs : {'अवधारणाएँ', 'नाम', 'और', 'मज़ा', 'सिखाना', 'पसंद', '।', 'इसमें', 'आता', 'है', 'एमएल', 'अवनीश', 'मुझे', 'कुछ', 'मेरा'}
No of Unique Words or Vocabs: 15
```

[ ]: