

```
In [1]: import nltk  
        from nltk.tokenize import word_tokenize
```

```
In [2]: p="Hello I am Ankit I am from Kanpur"
```

```
In [4]: print(word_tokenize(p))  
  
['Hello', 'I', 'am', 'Ankit', 'I', 'am', 'from', 'Kanpur']
```

```
In [5]: #POS tagging (basically tells the part of speech or each token information)  
        words=nltk.word_tokenize(p)  
        tagged_word=nltk.pos_tag(words)
```

```
In [6]: words
```

```
Out[6]: ['Hello', 'I', 'am', 'Ankit', 'I', 'am', 'from', 'Kanpur']
```

```
In [7]: tagged_word      #each token and information about it
```

```
Out[7]: [('Hello', 'NNP'),  
         ('I', 'PRP'),  
         ('am', 'VBP'),  
         ('Ankit', 'NNP'),  
         ('I', 'PRP'),  
         ('am', 'VBP'),  
         ('from', 'IN'),  
         ('Kanpur', 'NNP')]
```

```
In [8]: #PorterStemmer is a book for root words  
        from nltk.stem import PorterStemmer
```

```
In [9]: e=['go', 'went', 'gone']
```

```
In [10]: ps=PorterStemmer()
```

```
In [12]: for w in e:
         rootw=ps.stem(w)
         print(rootw)
```

gone

```
In [13]: e=['run','ran','running']
```

```
In [14]: for w in e:
         rootw=ps.stem(w)
         print(rootw)
```

run