

Extract_location_using_NLP

April 24, 2020

```
[1]: import nltk
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag
from nltk.chunk import ne_chunk
doc = '''
According to a media report, Mumbai has the highest density of cars in India.
→Pune is in second place.

The density of private cars in Mumbai has gone up by 18% in 2 years. There are
→510 cars per km of road as opposed to 430 cars per km in 2016. This is
→almost five times that of Delhi (108 cars per km). Despite having fewer cars
→than Delhi, Mumbai is more congested due to low road space. Mumbai has 2,000
→km of roads compared to the national capital, which has 28,000 km of
→roadways.

There are 10.2 lakh private cars in Mumbai. That is 28% of the total number of
→vehicles in the city, which stands at 36 lakh. According to RTO officials,
→the western suburbs have the highest number of registered cars (5 lakh).
→There are 3.3 lakh private cars in the island city and 1.7 lakh in the
→eastern suburbs.

Pune has 359 cars per km and Kolkata is the third most congested city with 319
→cars per km. Chennai comes in fourth with 297 cars per km followed by
→Bangalore with 149 cars per km.'''
word_tokens = nltk.word_tokenize(doc)
print(word_tokens)
```

```
['According', 'to', 'a', 'media', 'report', ',', ',', 'Mumbai', 'has', 'the',
'highest', 'density', 'of', 'cars', 'in', 'India', '.', 'Pune', 'is', 'in',
'second', 'place', '.', 'The', 'density', 'of', 'private', 'cars', 'in',
'Mumbai', 'has', 'gone', 'up', 'by', '18', '%', 'in', '2', 'years', '.',
'There', 'are', '510', 'cars', 'per', 'km', 'of', 'road', 'as', 'opposed', 'to',
'430', 'cars', 'per', 'km', 'in', '2016', '.', 'This', 'is', 'almost', 'five',
'times', 'that', 'of', 'Delhi', '(', '108', 'cars', 'per', 'km', ')', '.',
'Despite', 'having', 'fewer', 'cars', 'than', 'Delhi', ',', 'Mumbai', 'is',
'more', 'congested', 'due', 'to', 'low', 'road', 'space', '.', 'Mumbai', 'has',
```

'2,000', 'km', 'of', 'roads', 'compared', 'to', 'the', 'national', 'capital',
,', 'which', 'has', '28,000', 'km', 'of', 'roadways', '.', 'There', 'are',
'10.2', 'lakh', 'private', 'cars', 'in', 'Mumbai', '.', 'That', 'is', '28', '%',
'of', 'the', 'total', 'number', 'of', 'vehicles', 'in', 'the', 'city', ',',
'which', 'stands', 'at', '36', 'lakh', '.', 'According', 'to', 'RTO',
'officials', ',', 'the', 'western', 'suburbs', 'have', 'the', 'highest',
'number', 'of', 'registered', 'cars', '(', '5', 'lakh', ')', '.', 'There',
'are', '3.3', 'lakh', 'private', 'cars', 'in', 'the', 'island', 'city', 'and',
'1.7', 'lakh', 'in', 'the', 'eastern', 'suburbs', '.', 'Pune', 'has', '359',
'cars', 'per', 'km', 'and', 'Kolkata', 'is', 'the', 'third', 'most',
'congested', 'city', 'with', '319', 'cars', 'per', 'km', '.', 'Chennai',
'comes', 'in', 'fourth', 'with', '297', 'cars', 'per', 'km', 'followed', 'by',
'Bangalore', 'with', '149', 'cars', 'per', 'km', '.']

```
[2]: postags = nltk.pos_tag(word_tokens)
      print(postags)
```

[('According', 'VBG'), ('to', 'TO'), ('a', 'DT'), ('media', 'NNS'), ('report',
'NN'), (',', ','), ('Mumbai', 'NNP'), ('has', 'VBZ'), ('the', 'DT'), ('highest',
'JJ'), ('density', 'NN'), ('of', 'IN'), ('cars', 'NNS'), ('in', 'IN'),
(('India', 'NNP'), (',', ','), ('Pune', 'NNP'), ('is', 'VBZ'), ('in', 'IN'),
(('second', 'JJ'), ('place', 'NN'), (',', ','), ('The', 'DT'), ('density', 'NN'),
(('of', 'IN'), ('private', 'JJ'), ('cars', 'NNS'), ('in', 'IN'), ('Mumbai',
'NNP'), ('has', 'VBZ'), ('gone', 'VBN'), ('up', 'RP'), ('by', 'IN'), ('18',
'CD'), ('%', 'NN'), ('in', 'IN'), ('2', 'CD'), ('years', 'NNS'), (',', ','),
(('There', 'EX'), ('are', 'VBP'), ('510', 'CD'), ('cars', 'NNS'), ('per', 'IN'),
(('km', 'NN'), ('of', 'IN'), ('road', 'NN'), ('as', 'IN'), ('opposed', 'VBN'),
(('to', 'TO'), ('430', 'CD'), ('cars', 'NNS'), ('per', 'IN'), ('km', 'NN'),
(('in', 'IN'), ('2016', 'CD'), (',', ','), ('This', 'DT'), ('is', 'VBZ'),
(('almost', 'RB'), ('five', 'CD'), ('times', 'NNS'), ('that', 'IN'), ('of',
'IN'), ('Delhi', 'NNP'), (',', ','), ('108', 'CD'), ('cars', 'NNS'), ('per',
'IN'), ('km', 'NN'), (',', ','), (',', ','), ('Despite', 'IN'), ('having',
'VBG'), ('fewer', 'JJR'), ('cars', 'NNS'), ('than', 'IN'), ('Delhi', 'NNP'),
(',', ','), ('Mumbai', 'NNP'), ('is', 'VBZ'), ('more', 'RBR'), ('congested',
'JJ'), ('due', 'JJ'), ('to', 'TO'), ('low', 'JJ'), ('road', 'NN'), ('space',
'NN'), (',', ','), ('Mumbai', 'NNP'), ('has', 'VBZ'), ('2,000', 'CD'), ('km',
'NN'), ('of', 'IN'), ('roads', 'NNS'), ('compared', 'VBN'), ('to', 'TO'),
(('the', 'DT'), ('national', 'JJ'), ('capital', 'NN'), (',', ','), ('which',
'WDT'), ('has', 'VBZ'), ('28,000', 'CD'), ('km', 'NN'), ('of', 'IN'),
(('roadways', 'NNS'), (',', ','), ('There', 'EX'), ('are', 'VBP'), ('10.2',
'CD'), ('lakh', 'JJ'), ('private', 'JJ'), ('cars', 'NNS'), ('in', 'IN'),
(('Mumbai', 'NNP'), (',', ','), ('That', 'DT'), ('is', 'VBZ'), ('28', 'CD'),
(('%', 'NN'), ('of', 'IN'), ('the', 'DT'), ('total', 'JJ'), ('number', 'NN'),
(('of', 'IN'), ('vehicles', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('city', 'NN'),
(',', ','), ('which', 'WDT'), ('stands', 'VBZ'), ('at', 'IN'), ('36', 'CD'),
(('lakh', 'NN'), (',', ','), ('According', 'VBG'), ('to', 'TO'), ('RTO', 'NNP'),
(('officials', 'NNS'), (',', ','), ('the', 'DT'), ('western', 'JJ'), ('suburbs',

```
'NNS'), ('have', 'VBP'), ('the', 'DT'), ('highest', 'JJS'), ('number', 'NN'),
('of', 'IN'), ('registered', 'JJ'), ('cars', 'NNS'), ('(', '('), ('5', 'CD'),
('lakh', 'NN'), (')', ')'), ('.', '.'), ('There', 'EX'), ('are', 'VBP'), ('3.3',
'CD'), ('lakh', 'JJ'), ('private', 'JJ'), ('cars', 'NNS'), ('in', 'IN'), ('the',
'DT'), ('island', 'NN'), ('city', 'NN'), ('and', 'CC'), ('1.7', 'CD'), ('lakh',
'NN'), ('in', 'IN'), ('the', 'DT'), ('eastern', 'JJ'), ('suburbs', 'NN'), ('.',
'.'), ('Pune', 'NNP'), ('has', 'VBZ'), ('359', 'CD'), ('cars', 'NNS'), ('per',
'IN'), ('km', 'NN'), ('and', 'CC'), ('Kolkata', 'NNP'), ('is', 'VBZ'), ('the',
'DT'), ('third', 'JJ'), ('most', 'RBS'), ('congested', 'JJ'), ('city', 'NN'),
('with', 'IN'), ('319', 'CD'), ('cars', 'NNS'), ('per', 'IN'), ('km', 'NN'),
('.', '.'), ('Chennai', 'NNP'), ('comes', 'VBZ'), ('in', 'IN'), ('fourth',
'JJ'), ('with', 'IN'), ('297', 'CD'), ('cars', 'NNS'), ('per', 'IN'), ('km',
'NN'), ('followed', 'VBN'), ('by', 'IN'), ('Bangalore', 'NNP'), ('with', 'IN'),
('149', 'CD'), ('cars', 'NNS'), ('per', 'IN'), ('km', 'NN'), ('.', '.')]

```

```
[4]: nouns = ne_chunk(postags)
[5]: for x in str(nouns).split('\n'):
      if '/NN' in x:
          print(x)

```

```
media/NNS
report/NN
(PERSON Mumbai/NNP)
density/NN
cars/NNS
(GPE India/NNP)
Pune/NNP
place/NN
density/NN
cars/NNS
(GPE Mumbai/NNP)
%/NN
years/NNS
cars/NNS
km/NN
road/NN
cars/NNS
km/NN
times/NNS
(GPE Delhi/NNP)
cars/NNS
km/NN
cars/NNS
(PERSON Delhi/NNP)
(PERSON Mumbai/NNP)
road/NN
space/NN

```

(PERSON Mumbai/NNP)
km/NN
roads/NNS
capital/NN
km/NN
roadways/NNS
cars/NNS
(GPE Mumbai/NNP)
%/NN
number/NN
vehicles/NNS
city/NN
lakh/NN
(ORGANIZATION RTO/NNP)
officials/NNS
suburbs/NNS
number/NN
cars/NNS
lakh/NN
cars/NNS
island/NN
city/NN
lakh/NN
suburbs/NN
(PERSON Pune/NNP)
cars/NNS
km/NN
(PERSON Kolkata/NNP)
city/NN
cars/NNS
km/NN
(PERSON Chennai/NNP)
cars/NNS
km/NN
(ORGANIZATION Bangalore/NNP)
cars/NNS
km/NN

[]: