ca2-12214175

April 8, 2025

TEXT ANALYTICS-CA2 NAME:T.SAI KIRAN ROLL NO:14 REG:12214175 CSE(BIG DATA ANALYTICS)

[48]:
```
import nltk
from nltk.corpus import gazetteers

nltk.download('punkt')
nltk.download('gazetteers')

text = "i like  mostly India,Vietnam"

words = nltk.word_tokenize(text)

country = gazetteers.words('countries.txt')

countries = set(word for word in words if word in country)



print("Countries :", countries)
```

```
Countries : {'India', 'Vietnam'}
```

```
[nltk_data] Downloading package punkt to C:\Users\SAI
[nltk_data]     KIRAN\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package gazetteers to C:\Users\SAI
[nltk_data]     KIRAN\AppData\Roaming\nltk_data…
[nltk_data]   Package gazetteers is already up-to-date!
```

[26]:
```
import nltk
text = "i am a good fellow playing in ground and is he good"
words = nltk.word_tokenize(text)
pos_tag=nltk.pos_tag(words)
nouns=[word for word,tag in pos_tag if tag in ['NN']]
verbs=[word for word,tag in pos_tag if tag in ['VBZ',"VBP"]]
print("pos tag: ",pos_tag)
print()
print("nouns:",nouns)
```

```
print()
print("verbs:",verbs)
```

```
pos tag:  [('i', 'NN'), ('am', 'VBP'), ('a', 'DT'), ('good', 'JJ'), ('fellow',
'NN'), ('playing', 'NN'), ('in', 'IN'), ('ground', 'NN'), ('and', 'CC'), ('is',
'VBZ'), ('he', 'PRP'), ('good', 'JJ')]

nouns: ['i', 'fellow', 'playing', 'ground']

verbs: ['am', 'is']
```

[ ]: