In [1]:

```
text = "My name is harshita khangarot, I m asst. Prof. in JIET-jodhpur. I have 7 Ye
```

In [2]:

```
from nltk.tokenize import WordPunctTokenizer
tokenizer = WordPunctTokenizer()
token=tokenizer.tokenize(text)
```

In [3]:

```
token
```

Out[3]:

```
['My',
 'name',
 'is',
 'harshita',
 'khangarot',
 ',',
 'I',
 'm',
 'asst',
 '.',
 'Prof',
 '.',
 'in',
 'JIET',
 '-',
 'jodhpur',
 '.',
 'I',
 'have',
 '7',
 'Years',
 'of',
 'Exp',
 ',',
 'I',
 'love',
 'Jodhpur',
 'and',
 'Jodhpur',
 "'",
 's',
 'sweets',
 '.']
```

In [5]:

```
words=[]
for word in token:
    words.append(word.lower())
words[:]
```

Out[5]:
```
['my',
 'name',
 'is',
 'harshita',
 'khangarot',
 ',',
 'i',
 'm',
 'asst',
 '.',
 'prof',
 '.',
 'in',
 'jiet',
 '-',
 'jodhpur',
 '.',
 'i',
 'have',
 '7',
 'years',
 'of',
 'exp',
 ',',
 'i',
 'love',
 'jodhpur',
 'and',
 'jodhpur',
 "'",
 's',
 'sweets',
 '.']
```

In [6]:
```
import nltk
nltk.download("stopwords")
```

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

Out[6]:
```
True
```

In [7]:
```
sw=nltk.corpus.stopwords.words('english')
sw[:5]
```

Out[7]:
```
['i', 'me', 'my', 'myself', 'we']
```

In [8]:
```
# get the list without stop words
words_ne=[]
for word in words:
    if word not in sw:
        words_ne.append(word)
words_ne[:]
```

Out[8]:

```
['name',
 'harshita',
 'khangarot',
 ',',
 'asst',
 '.',
 'prof',
 '.',
 'jiet',
 '-',
 'jodhpur',
 '.',
 '7',
 'years',
 'exp',
 ',',
 'love',
 'jodhpur',
 'jodhpur',
 "'",
 'sweets',
 '.']
```

In [9]:

```
freq = nltk.FreqDist(words)
freq
```

Out[9]:

```
FreqDist({'.': 4, 'i': 3, 'jodhpur': 3, ',': 2, 'my': 1, 'name': 1, 'is': 1, 'hars
hita': 1, 'khangarot': 1, 'm': 1, ...})
```

In [10]:

```
freq = nltk.FreqDist(words_ne)
freq
```

Out[10]:

```
FreqDist({'.': 4, 'jodhpur': 3, ',': 2, 'name': 1, 'harshita': 1, 'khangarot': 1,
'asst': 1, 'prof': 1, 'jiet': 1, '-': 1, ...})
```

In [11]:

```
nltk.download('punkt')
```

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

Out[11]:

```
True
```

In [12]:

```
from nltk.corpus import webtext
from nltk.probability import FreqDist
```

Q: Write a Program to Find the word frequency in any corpus and find only words they have length >10 and print in sorted order.

In [13]:

```
nltk.download('webtext')
wt_words = webtext.words("singles.txt")
data = nltk.FreqDist(wt_words)
# Let's take the specific words only if their frequency is greater than 10.
filter_words = dict([(m, n) for m, n in data.items() if len(m) > 10])
for key in sorted(filter_words):
    print("%s: %s" % (key, filter_words[key]))
data= nltk.FreqDist(filter_words)
print(data)
```

```
ABBREVIATIONS: 1
ADVENTUROUS: 1
AFFECTIONATE: 2
BUSINESSMAN: 4
Businessman: 1
DISCIPLINARIAN: 1
INTELLIGENT: 1
Intelligent: 1
Nationality: 1
PROFESSIONAL: 1
adventurous: 2
affectionate: 2
appreciated: 1
candlelight: 1
comfortable: 2
companionship: 2
complications: 1
conversation: 3
disappointed: 1
emotionally: 1
environment: 1
established: 1
financially: 4
fulfillment: 1
independance: 1
independant: 2
intelligent: 3
interesting: 1
judgemental: 1
permaculture: 1
personality: 3
professional: 4
relationship: 29
responsible: 2
sufficiencies: 1
trustworthy: 1
understanding: 1
unimportant: 1
<FreqDist with 38 samples and 88 outcomes>
```

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

```
In [ ]:
```