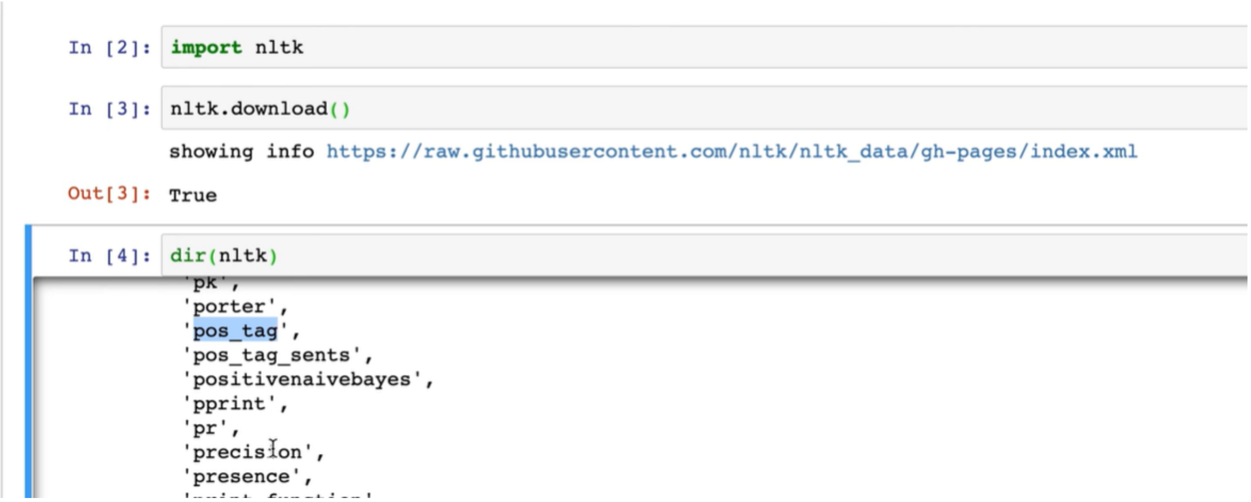
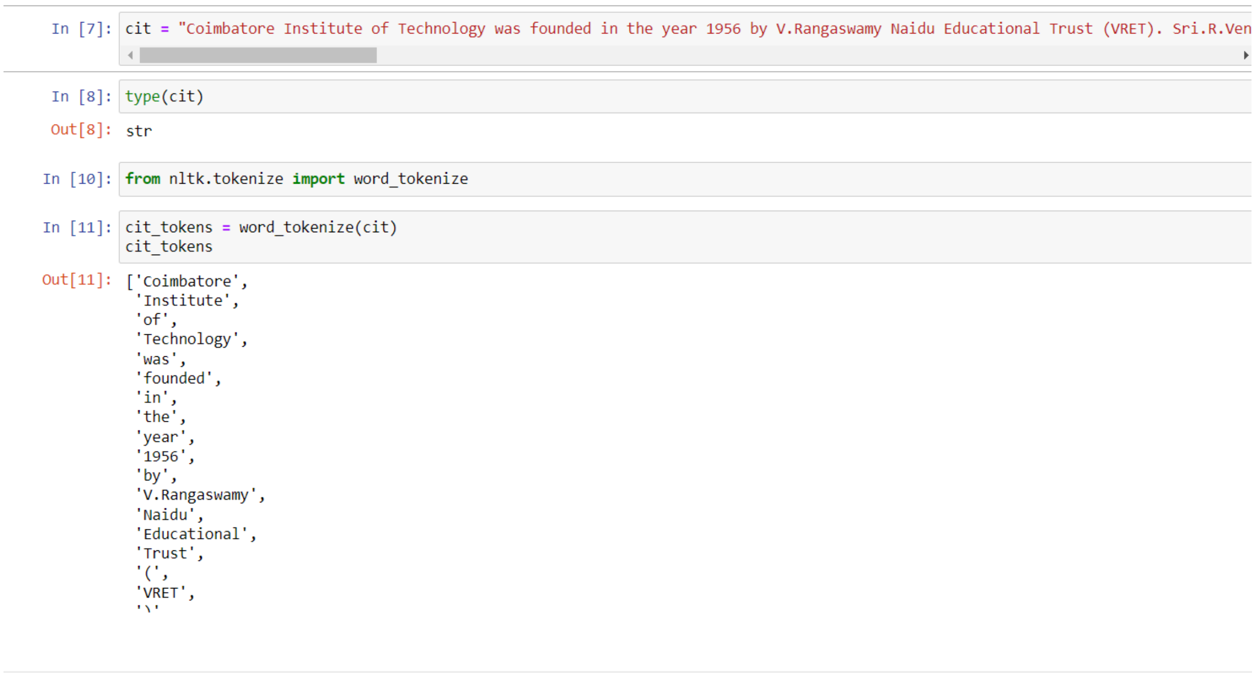
1.Import NLTK package-(Jupyter)
(a)Explore all the packages list in UI and
(b) use dir function to list all the functions
 (c) write any five functions in the observation note.



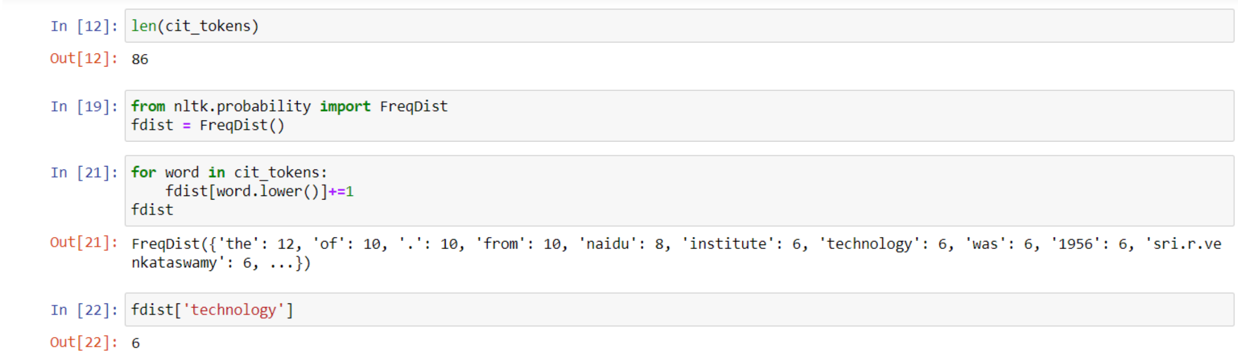
```
In [2]: import nltk

In [3]: nltk.download()
        showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
Out[3]: True

In [4]: dir(nltk)
         'pk',
         'porter',
         'pos_tag',
         'pos_tag_sents',
         'positivenaivebayes',
         'pprint',
         'pr',
         'precision',
         'presence',
         'print_function'
```

2..Implement a program to using word tokenizer

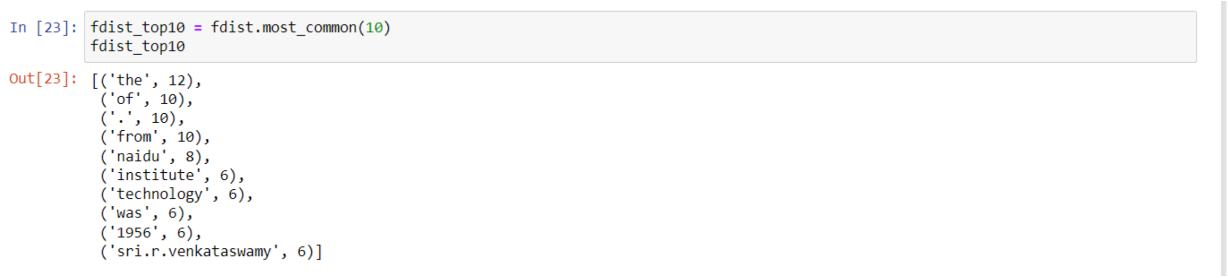
1. Create your own corpus for about 500 words
2. Find the length of tokens
3. Find the number of sentences in it
4. Use frequency distribution function to find the occurrence of words
5. Find the occurrence of particular word
6. Top 5 highest frequency of words in the document
7. Show a dispersion plot for the above(matplotlib/ or any)



```
In [7]:  cit = "Coimbatore Institute of Technology was founded in the year 1956 by V.Rangaswamy Naidu Educational Trust (VRET). Sri.R.Ven

In [8]:  type(cit)

Out[8]:  str

In [10]: from nltk.tokenize import word_tokenize

In [11]: cit_tokens = word_tokenize(cit)
         cit_tokens

Out[11]: ['Coimbatore',
          'Institute',
          'of',
          'Technology',
          'was',
          'founded',
          'in',
          'the',
          'year',
          '1956',
          'by',
          'V.Rangaswamy',
          'Naidu',
          'Educational',
          'Trust',
          '(',
          'VRET',
          ')'
```



```
In [12]: len(cit_tokens)

Out[12]: 86

In [19]: from nltk.probability import FreqDist
         fdist = FreqDist()

In [21]: for word in cit_tokens:
             fdist[word.lower()]+=1
         fdist

Out[21]: FreqDist({'the': 12, 'of': 10, '.': 10, 'from': 10, 'naidu': 8, 'institute': 6, 'technology': 6, 'was': 6, '1956': 6, 'sri.r.ve
         nkataswamy': 6, ...})

In [22]: fdist['technology']

Out[22]: 6
```



```
In [23]: fdist_top10 = fdist.most_common(10)
         fdist_top10

Out[23]: [('the', 12),
          ('of', 10),
          ('.', 10),
          ('from', 10),
          ('naidu', 8),
          ('institute', 6),
          ('technology', 6),
          ('was', 6),
          ('1956', 6),
          ('sri.r.venkataswamy', 6)]
```