# Winter Semester 2022 – 2023

# CSE4022 – Natural Language Processing

# Task – 2

Name: T Harshitha Deepthi

Reg no: 20BCE2019

Slot: E1+TE1

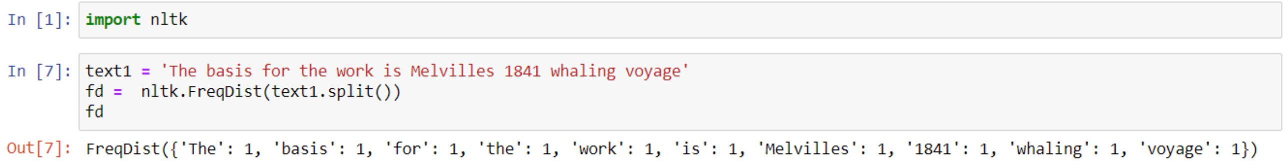
Venue: SJT626

## 1. Frequency distribution function of words in a text

text1 = ‘the basis for the work is Melvilles 1841 whaling voyage’

fd =  nltk.FreqDist(text1.split())

fd



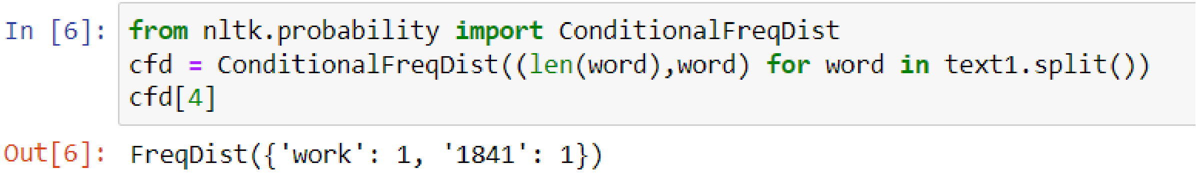
```
In [1]: import nltk

In [7]: text1 = 'The basis for the work is Melvilles 1841 whaling voyage'
        fd =  nltk.FreqDist(text1.split())
        fd

Out[7]: FreqDist({'The': 1, 'basis': 1, 'for': 1, 'the': 1, 'work': 1, 'is': 1, 'Melvilles': 1, '1841': 1, 'whaling': 1, 'voyage': 1})
```

## 2. conditional distribution

from nltk.probablity import ConditionalFreqDist

cfd = ConditionalFreqDist((len(word) for word in text1.split())

cfd[4]



```
In [6]: from nltk.probability import ConditionalFreqDist
        cfd = ConditionalFreqDist((len(word),word) for word in text1.split())
        cfd[4]

Out[6]: FreqDist({'work': 1, '1841': 1})
```

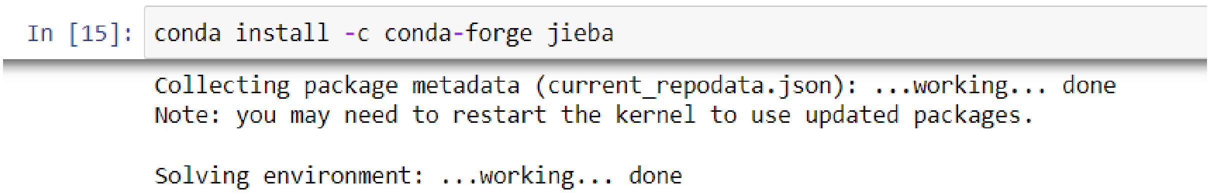
## 3. Chinese segmentation using jieba

*/ install jieba

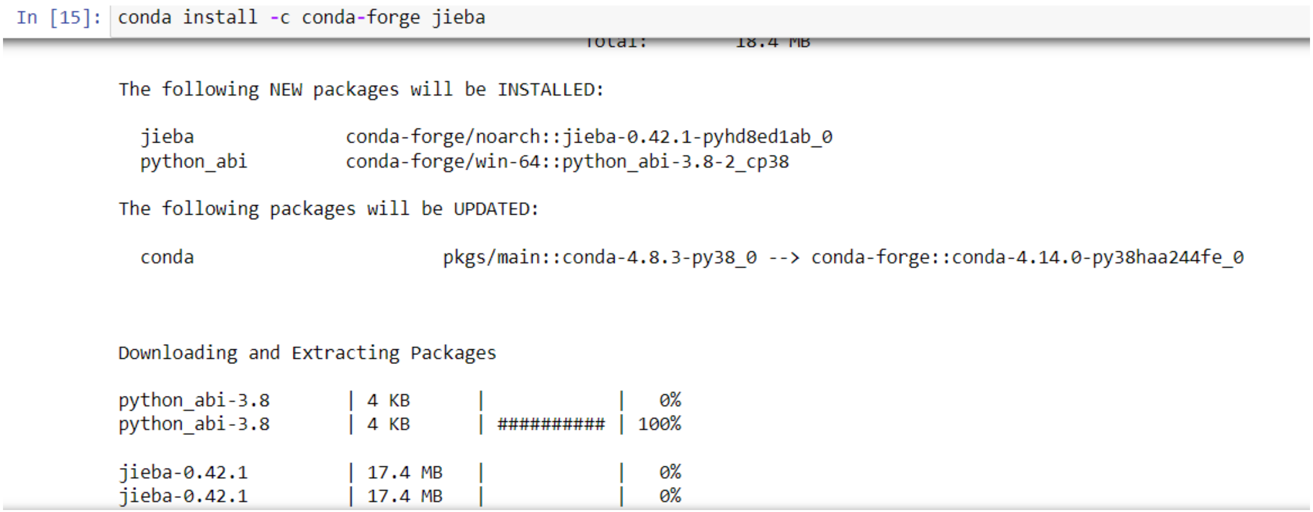
import jieba

seg = jieba.cut(“Chinese characters”,cut_all=True)

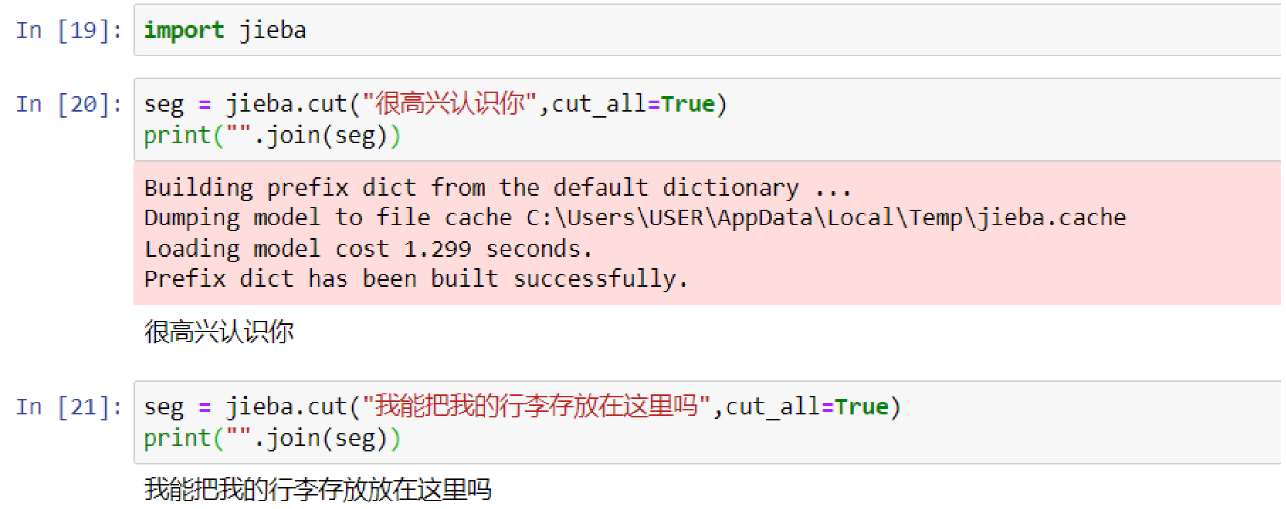
print(“ “.join(seg))



```
In [15]: conda install -c conda-forge jieba

         Collecting package metadata (current_repodata.json): ...working... done
         Note: you may need to restart the kernel to use updated packages.

         Solving environment: ...working... done
```



```
In [15]: conda install -c conda-forge jieba
                                        Total:          18.4 MB

         The following NEW packages will be INSTALLED:

           jieba              conda-forge/noarch::jieba-0.42.1-pyhd8ed1ab_0
           python_abi         conda-forge/win-64::python_abi-3.8-2_cp38

         The following packages will be UPDATED:

           conda                        pkgs/main::conda-4.8.3-py38_0 --> conda-forge::conda-4.14.0-py38haa244fe_0


         Downloading and Extracting Packages

         python_abi-3.8       | 4 KB      |            |   0%
         python_abi-3.8       | 4 KB      | ########## | 100%

         jieba-0.42.1         | 17.4 MB   |            |   0%
         jieba-0.42.1         | 17.4 MB   |            |   0%
```



```
In [19]: import jieba

In [20]: seg = jieba.cut("很高兴认识你",cut_all=True)
         print("".join(seg))

         Building prefix dict from the default dictionary ...
         Dumping model to file cache C:\Users\USER\AppData\Local\Temp\jieba.cache
         Loading model cost 1.299 seconds.
         Prefix dict has been built successfully.

         很高兴认识你

In [21]: seg = jieba.cut("我能把我的行李存放在这里吗",cut_all=True)
         print("".join(seg))

         我能把我的行李存放放在这里吗
```

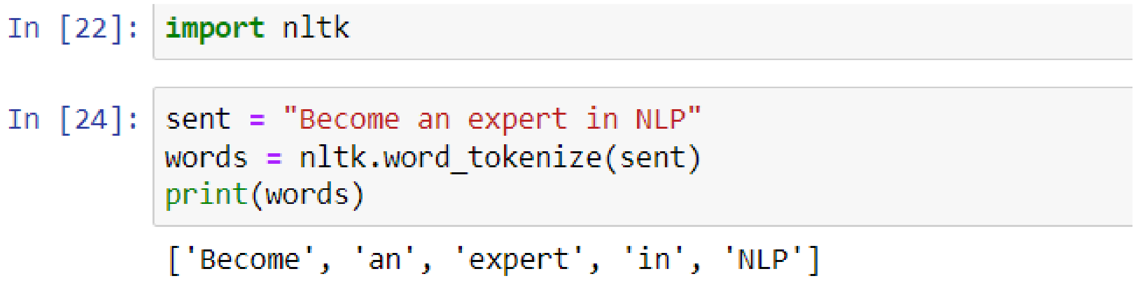
## 4. Printing Words

import nltk

sent = “Become an expert in NLP”

words = nltk.word_tokenize(sent)

print(words)

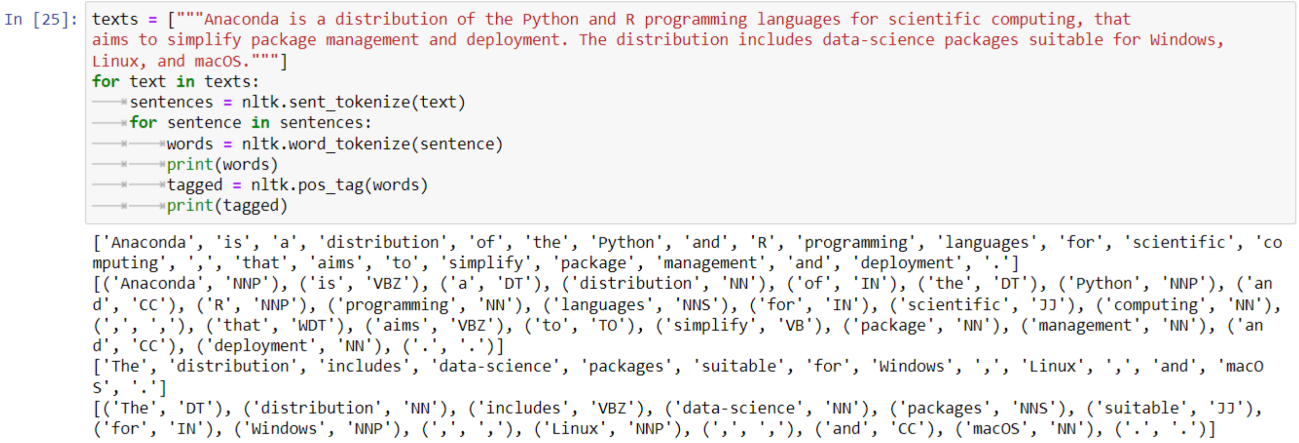


```
In [22]: import nltk
```

```
In [24]: sent = "Become an expert in NLP"
         words = nltk.word_tokenize(sent)
         print(words)

         ['Become', 'an', 'expert', 'in', 'NLP']
```

## 5. Printing tagged sentences

```
texts = [”””Anaconda is a distribution of the Python and R programming
languages for scientific computing, that aims to simplify package management
and deployment. The distribution includes data-science packages suitable for
Windows, Linux, and macOS.”””]

for text in texts:

    sentences = nltk.sent_tokenize(text)

    for sentence in sentences:

        words = nltk.word_tokenize(sentence)

        # print(words)

        # tagged = nltk.pos_tag(words)

        # print(tagged)
```



```
In [25]: texts = ["""Anaconda is a distribution of the Python and R programming languages for scientific computing, that
         aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows,
         Linux, and macOS."""]
         for text in texts:
             sentences = nltk.sent_tokenize(text)
             for sentence in sentences:
                 words = nltk.word_tokenize(sentence)
                 print(words)
                 tagged = nltk.pos_tag(words)
                 print(tagged)

         ['Anaconda', 'is', 'a', 'distribution', 'of', 'the', 'Python', 'and', 'R', 'programming', 'languages', 'for', 'scientific', 'co
         mputing', ',', 'that', 'aims', 'to', 'simplify', 'package', 'management', 'and', 'deployment', '.']
         [('Anaconda', 'NNP'), ('is', 'VBZ'), ('a', 'DT'), ('distribution', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Python', 'NNP'), ('an
         d', 'CC'), ('R', 'NNP'), ('programming', 'NN'), ('languages', 'NNS'), ('for', 'IN'), ('scientific', 'JJ'), ('computing', 'NN'),
         (',', ','), ('that', 'WDT'), ('aims', 'VBZ'), ('to', 'TO'), ('simplify', 'VB'), ('package', 'NN'), ('management', 'NN'), ('an
         d', 'CC'), ('deployment', 'NN'), ('.', '.')]
         ['The', 'distribution', 'includes', 'data-science', 'packages', 'suitable', 'for', 'Windows', ',', 'Linux', ',', 'and', 'macO
         S', '.']
         [('The', 'DT'), ('distribution', 'NN'), ('includes', 'VBZ'), ('data-science', 'NN'), ('packages', 'NNS'), ('suitable', 'JJ'),
         ('for', 'IN'), ('Windows', 'NNP'), (',', ','), ('Linux', 'NNP'), (',', ','), ('and', 'CC'), ('macOS', 'NN'), ('.', '.')]
```