21MIC0087 Hands-on-session 2

```
In [1]: import nltk
```

```
In [28]: text1="Natural language processing (NLP) refers to the branch of computer sc
```

```
In [3]: fd=nltk.FreqDist(text1.split())
```

```
In [5]: text1.split()
```

```
Out[5]: ['Natural',
         'language',
         'processing',
         '(NLP)',
         'refers',
         'to',
         'the',
         'branch',
         'of',
         'computer',
         'science—and',
         'more',
         'specifically,',
         'the',
         'branch',
         'of',
         'artificial',
         'intelligence',
         'or',
         'AI—concerned',
         'with',
         'giving',
         'computers',
         'the',
         'ability',
         'to',
         'understand',
         'text',
         'and',
         'spoken',
         'words',
         'in',
         'much',
         'the',
         'same',
         'way',
         'human',
         'beings',
         'can.']
```

```
In [6]: fd
```

```
Out[6]: FreqDist({'the': 4, 'to': 2, 'branch': 2, 'of': 2, 'Natural': 1, 'languag
        e': 1, 'processing': 1, '(NLP)': 1, 'refers': 1, 'computer': 1, ...})
```

```
In [9]: from nltk.corpus import inaugural
```

```
In [23]: text2=inaugural.words(fileids='1861-Lincoln.txt')[1000:2000]
```

In [24]:

```
nltk.FreqDist(text2)
```

Out[24]:

```
FreqDist({'the': 70, ',': 51, 'of': 38, 'to': 32, 'and': 30, '.': 24, 'be':
23, 'in': 19, 'Union': 16, 'it': 16, ...})
```

In [31]:

```
from nltk.probability import ConditionalFreqDist
cfd=ConditionalFreqDist((len(word),word)for word in text2)
```

In [39]:

```
cfd[13]
```

Out[39]:

```
FreqDist({'contemplation': 2, 'circumstances': 2, 'Confederation': 1, 'revo
lutionary': 1, 'authoritative': 1, 'impracticable': 1})
```

In [ ]:

```

```

In [40]:

```
pip install jieba
```

```
Collecting jieba
  Downloading jieba-0.42.1.tar.gz (19.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.2/19.2 MB 279.3 kB/s eta 0:
00:00m eta 0:00:01[36m0:00:02
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: jieba
  Building wheel for jieba (setup.py) ... done
  Created wheel for jieba: filename=jieba-0.42.1-py3-none-any.whl size=1931
4458 sha256=e5cf219d08580be4afda223c1c07c8e894b412ebf4d8cf7db3548b62cb7cc9b
1
  Stored in directory: /home/vedant/.cache/pip/wheels/ac/60/cf/538a1f183409
caf1fc136b5d2c2dee329001ef6da2c5084bef
Successfully built jieba
Installing collected packages: jieba
Successfully installed jieba-0.42.1
Note: you may need to restart the kernel to use updated packages.
```

In [41]:

```
import jieba
```

In [42]:

```
seg_list=jieba.cut("你好吗",cut_all=True)
```

In [46]:

```
print(seg_list)
```

```
<generator object Tokenizer.cut at 0x7f7674ce62a0>
```

In [47]:

```
print(",".join(seg_list))
```

```
Building prefix dict from the default dictionary ...
Dumping model to file cache /tmp/jieba.cache
Loading model cost 0.779 seconds.
Prefix dict has been built successfully.
你好,吗
```

Splitted based on greedy segmentation algorithm

Manual word phoenitic

mæn.ju.əl

In [ ]:

```

```