Module 2 (Python 3)

Basic NLP Tasks with NLTK

In [1]:

```
import nltk
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Counting vocabulary of words

In [5]:

```
texts()
```

```
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

In [4]:

```
sents()
```

```
sent1: Call me Ishmael .
sent2: The family of Dashwood had long been settled in Sussex .
sent3: In the beginning God created the heaven and the earth .
sent4: Fellow - Citizens of the Senate and of the House of Representatives :
sent5: I have a problem with people PMing me to lol JOIN
sent6: SCENE 1 : [ wind ] [ clop clop clop ] KING ARTHUR : Whoa there !
sent7: Pierre Vinken , 61 years old , will join the board as a nonexecutive
director Nov. 29 .
sent8: 25 SEXY MALE , seeks attrac older single lady , for discreet encounte
rs .
sent9: THE suburb of Saffron Park lay on the sunset side of London , as red
and ragged as a cloud of sunset .
```

In [2]:

```
text7
```

Out[2]:

```
<Text: Wall Street Journal>
```

In [3]:

```
sent7
```

Out[3]:

```
['Pierre',
 'Vinken',
 ',',
 '61',
 'years',
 'old',
 ',',
 'will',
 'join',
 'the',
 'board',
 'as',
 'a',
 'nonexecutive',
 'director',
 'Nov.',
 '29',
 '.']
```

In [6]:

```
len(sent7)
```

Out[6]:

```
18
```

In [7]:

```
len(text7)
```

Out[7]:

```
100676
```

In [8]:

```
len(set(text7))
```

Out[8]:

```
12408
```

In [13]:

```
list(set(text7))[:10]
```

Out[13]:

```
['Blanchard',
 '1.56',
 'director',
 'Continental',
 'logistical',
 'generation',
 'jugglers',
 'reclaimed',
 'TREASURY',
 'might']
```

Frequency of words

In [10]:

```
dist = FreqDist(text7)
len(dist)
```

Out[10]:

```
12408
```

In [12]:

```
vocab1 = dist.keys()
#vocab1[:10]
# In Python 3 dict.keys() returns an iterable view instead of a list
list(vocab1)[:10]
```

Out[12]:

```
['Pierre', 'Vinken', ',', '61', 'years', 'old', 'will', 'join', 'the', 'boar
d']
```

In [14]:

```
dist['four']
```

Out[14]:

```
20
```

In [15]:

```
freqwords = [w for w in vocab1 if len(w) > 5 and dist[w] > 100]
freqwords
```

Out[15]:

```
['billion',
 'company',
 'president',
 'because',
 'market',
 'million',
 'shares',
 'trading',
 'program']
```

Normalization and stemming

In [20]:

```
input1 = "List listed lists listing listings"
words1 = input1.lower().split(' ')               # Normalization
words1
```

Out[20]:

```
['list', 'listed', 'lists', 'listing', 'listings']
```

In [21]:

```
porter = nltk.PorterStemmer()
[porter.stem(t) for t in words1]                  # Stemming
```

Out[21]:

```
['list', 'list', 'list', 'list', 'list']
```

Lemmatization

In [22]:

```
udhr = nltk.corpus.udhr.words('English-Latin1')    # udhr - universal declaration of human
udhr[:20]
```

Out[22]:

```
['Universal',
 'Declaration',
 'of',
 'Human',
 'Rights',
 'Preamble',
 'Whereas',
 'recognition',
 'of',
 'the',
 'inherent',
 'dignity',
 'and',
 'of',
 'the',
 'equal',
 'and',
 'inalienable',
 'rights',
 'of']
```

In [23]:

```
[porter.stem(t) for t in udhr[:20]] # Still Lemmatization
```

Out[23]:

```
['univers',
 'declar',
 'of',
 'human',
 'right',
 'preambl',
 'wherea',
 'recognit',
 'of',
 'the',
 'inher',
 'digniti',
 'and',
 'of',
 'the',
 'equal',
 'and',
 'inalien',
 'right',
 'of']
```

In [25]:

```
WNlemma = nltk.WordNetLemmatizer()
[WNlemma.lemmatize(t) for t in udhr[:20]]    # check udhr[4] and udhr[18]
```

Out[25]:

```
['Universal',
 'Declaration',
 'of',
 'Human',
 'Rights',
 'Preamble',
 'Whereas',
 'recognition',
 'of',
 'the',
 'inherent',
 'dignity',
 'and',
 'of',
 'the',
 'equal',
 'and',
 'inalienable',
 'right',
 'of']
```

Tokenization

In [26]:

```
text11 = "Children shouldn't drink a sugary drink before bed."
text11.split(' ')
```

Out[26]:

```
['Children', "shouldn't", 'drink', 'a', 'sugary', 'drink', 'before', 'bed.']
```

In [27]:

```
nltk.word_tokenize(text11)
```

Out[27]:

```
['Children',
 'should',
 "n't",
 'drink',
 'a',
 'sugary',
 'drink',
 'before',
 'bed',
 '.']
```

In [28]:

```
text12 = "This is the first sentence. A gallon of milk in the U.S. costs $2.99. Is this the
sentences = nltk.sent_tokenize(text12)
len(sentences)
```

Out[28]:

```
4
```

In [29]:

```
sentences
```

Out[29]:

```
['This is the first sentence.',
 'A gallon of milk in the U.S. costs $2.99.',
 'Is this the third sentence?',
 'Yes, it is!']
```

Advanced NLP Tasks with NLTK

POS tagging

In [30]:

```
nltk.help.upenn_tagset('MD')  # 'CC', 'CD', 'DT', 'IN', 'JJ', 'NN', 'POS', 'PRP', 'RB', 'SY
```

```
MD: modal auxiliary
    can cannot could couldn't dare may might must need ought shall should
    shouldn't will would
```

In [31]:

```
text13 = nltk.word_tokenize(text11)
nltk.pos_tag(text13)
```

Out[31]:

```
[('Children', 'NNP'),
 ('should', 'MD'),
 ("n't", 'RB'),
 ('drink', 'VB'),
 ('a', 'DT'),
 ('sugary', 'JJ'),
 ('drink', 'NN'),
 ('before', 'IN'),
 ('bed', 'NN'),
 ('.', '.')]
```