D11 詞性標註(Part-of-speech tagging)

重要知識點

- 了解如何使用Jieba/nltk完成中英文詞性標注

資料來源：http://parts-of-speech.info/

中文斷詞器：結巴 / Chinese word Segmentation: Jieba

Jieba 是目前中文界最流行的斷詞演算法，並且已在 Github 上開源。

NLTK ：Natural Language Tool Kit

nltk 是一套基於 Python 的自然語言處理工具箱，提供多個自然語言相關資料集與 processing API。

基本安裝

- 安裝 jibe
  pip install jieba (pip3 install jieba)
- 安裝 nltk
  pip install nltk (pip3 install nltk)

Jieba/NLTK 詞性標註

- 今天課程為程式練習，學員們可以直接操作 ipynb 檔。
- 程式碼部分都有註解，大家也可以自己調整參數，或輸入不同字串練習。

資料來源：結巴斷詞器與FastTag

jieba 斷句


```python
sentence_1 = "我愛寫程式"
sentence_2 = "Python是一種廣泛使用的直譯式、進階程式、通用型程式語言"
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
print("output 全模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=True, HMM=False))))
```

jieba 新增單詞


```python
sentence_1 = "精通各種程式語言是一件相當困難的事情"
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
jieba.add_word('程式語言')
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
```

讀入字典新增單詞


```python
## 新增單詞 ，格式：每行包含一個單詞 詞頻(可省略) 詞性(可省略)
new_words = '程式語言\nCupoy平台\n自然語言處理'
with open('new_words.txt', 'w') as file:
  file.write(new_words)
sentence_1 = "我在Cupoy平台上學習自然語言處理"
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
## 讀入字典
jieba.load_userdict('new_words.txt')
sentence_1 = "我在Cupoy平台上學習自然語言處理"
print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False))))
```

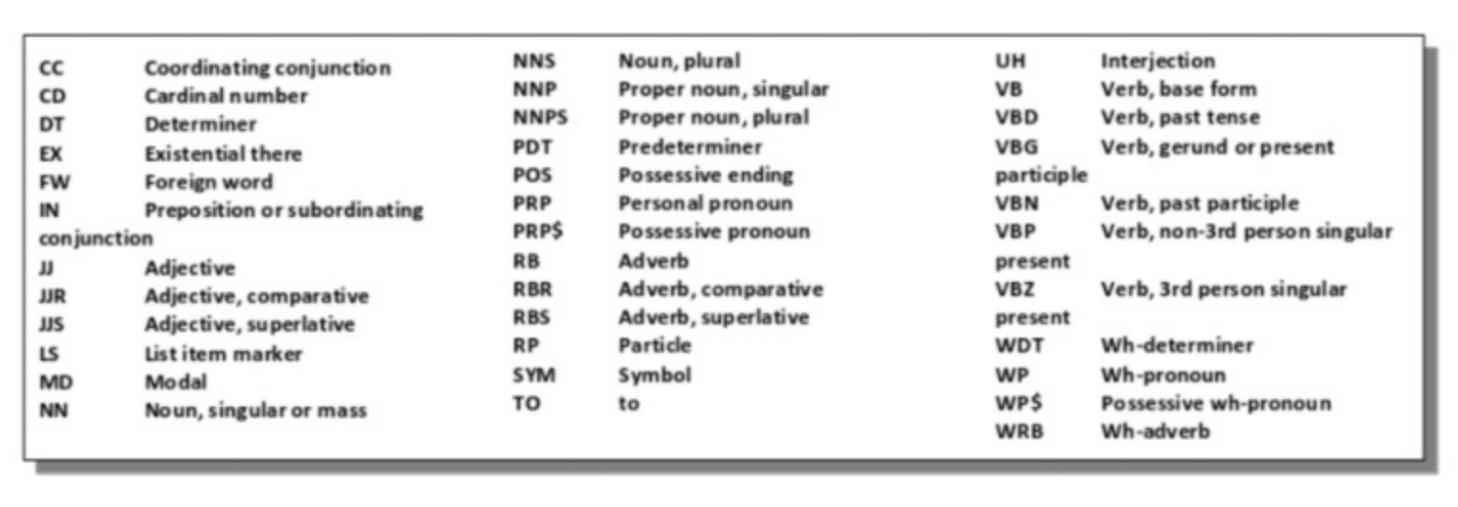
Tokenize：可以用來取出斷詞位置


```python
sentence_2 = "Python是一個廣泛使用的直譯式、進階程式、通用型程式語言"
words = jieba.tokenize(sentence_2,)
for tk in words:
  print("word: {},        start:{}, end:{}".format(tk[0],tk[1],tk[2]))
```

Pos Tagging

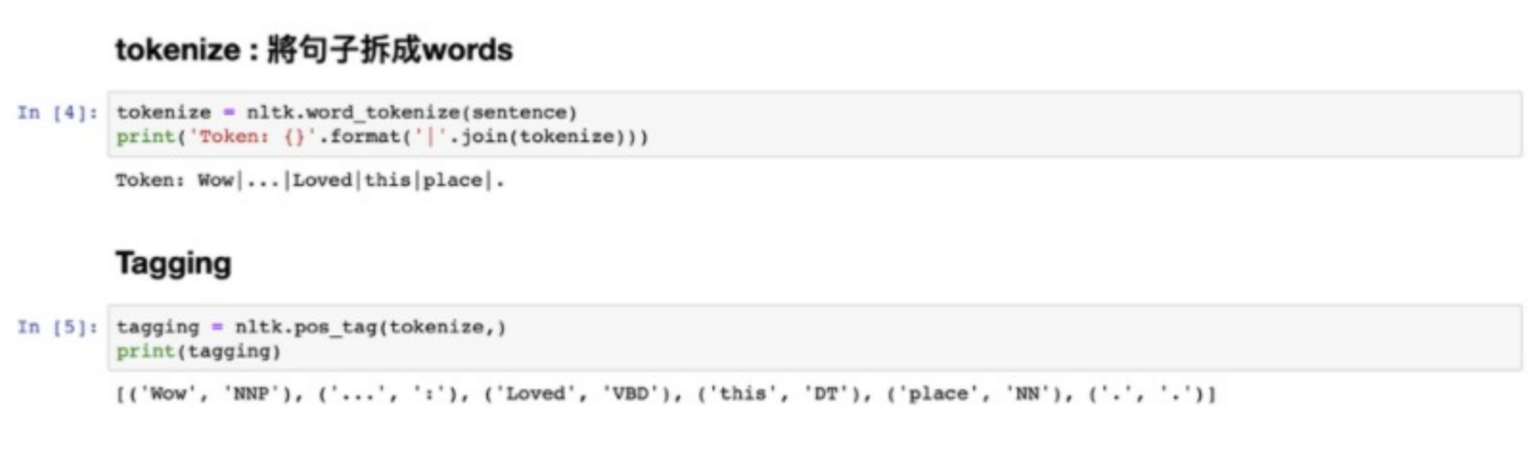

```python
import jieba.posseg as pseg
sentence_2 = "Python是一種廣泛使用的直譯式、進階程式、通用型程式語言"
words = pseg.cut(sentence_2,)
## 對應詞性可在此網頁查詢：https://www.cnblogs.com/chenbjin/p/4341930.html
for word, flag in words:
  print(word, flag)
```

nltk 詞性對照表



CC	Coordinating conjunction	NNS	Noun, plural	UH	Interjection
CD	Cardinal number	NNP	Proper noun, singular	VB	Verb, base form
DT	Determiner	NNPS	Proper noun, plural	VBD	Verb, past tense
EX	Existential there	PDT	Predeterminer	VBG	Verb, gerund or present participle
FW	Foreign word	POS	Possessive ending	VBN	Verb, past participle
IN	Preposition or subordinating conjunction	PRP	Personal pronoun	VBP	Verb, non-3rd person singular present
JJ	Adjective	PRP$	Possessive pronoun	VBZ	Verb, 3rd person singular present
JJR	Adjective, comparative	RB	Adverb	WDT	Wh-determiner
JJS	Adjective, superlative	RBR	Adverb, comparative	WP	Wh-pronoun
LS	List item marker	RBS	Adverb, superlative	WP$	Possessive wh-pronoun
MD	Modal	RP	Particle	WRB	Wh-adverb
NN	Noun, singular or mass	SYM	Symbol		
		TO	to		

tokenize / Tagging




```python
tokenize = nltk.word_tokenize(sentence)
print('Token: {}'.format('|'.join(tokenize)))
tagging = nltk.pos_tag(tokenize,)
print(tagging)
```

參考資料

網站：Jieba: Github

下一步：閱讀範例與完成作業