D06 使用結巴進行中文斷詞

簡報閱讀 範例與作業 問題討論

重要知識點
中文斷詞利器: 結巴(Jieba)
系統環境安裝
以jieba進行斷詞

重要知識點

- 如何使用結巴來進行實際的斷詞操作。

補充資料：簡單易用的中英文斷詞和詞性標註

中文斷詞利器: 結巴(Jieba)

- 要對非結構化的文字資料進行分析，第一件事情是對文字資料抽取結構化的量化數值特徵。除了用「文字探勘分析器」簡單分析字數、句數之外，最常見的分析方式就是斷詞和詞性分析。
- 作為最廣為人使用的斷詞器，結巴(Jieba)不只可搭配「非結構化資料分析：文本分類」等機器學習來使用，更可以用在質性研究的內容分析、文本分析或敘說分析上，先用Jieba找出特定詞性的文本內容來分析。

系統環境安裝

- 安裝Jieba, 只需在終端機上使用

```
pip install jieba
```

- 在.py的腳本中, 引入Jieba模組

```python
import jieba
```

以jieba進行斷詞

- 以結巴進行斷詞
- 結巴斷詞的輸出為generator, 需要使用遞迴將值取出

```python
import jieba

input_string = '小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造'
cutted_words = jieba.cut(input_string)
words = [word for word in cutted_words]
print(words)
```

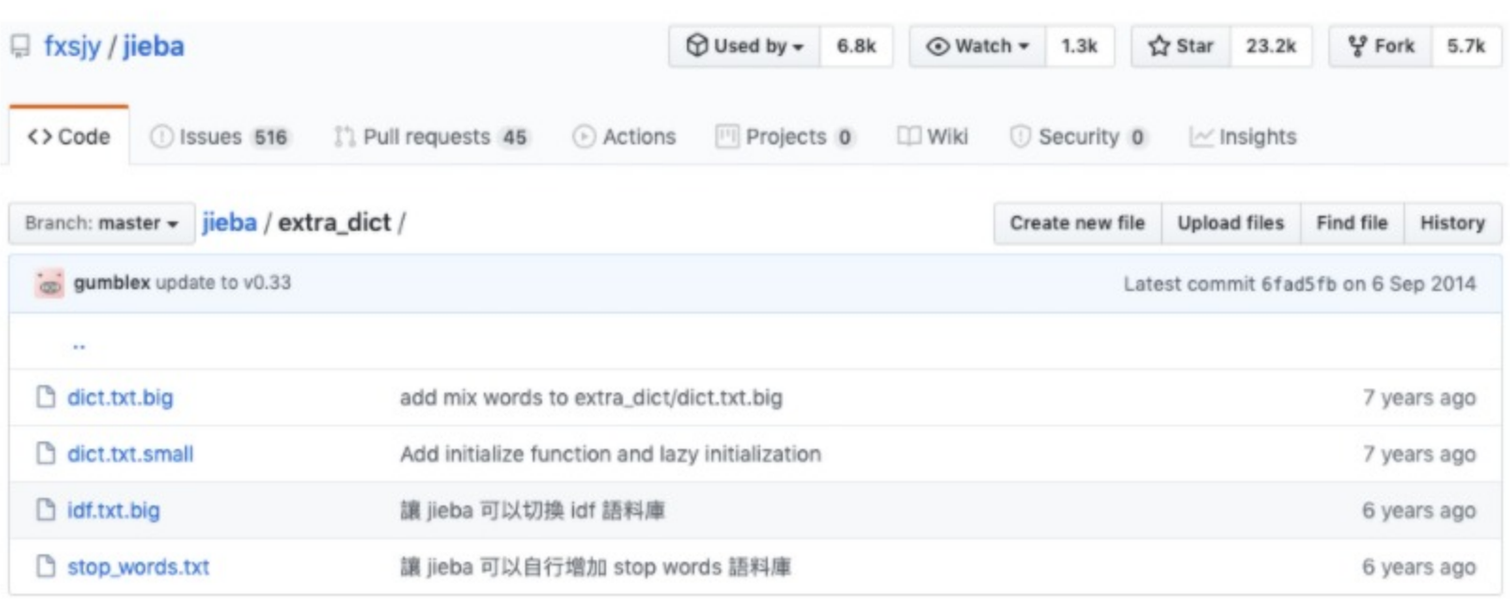
```
['小明', '碩士', '畢業', '於', '國立', '臺', '灣大學', '，', '現在', '在', '日本', '東京大學', '進修', '深造']
```

```python
import jieba
import jieba.posseg as pseg

input_str = '小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造'
cutted_words = pseg.cut(input_str)
words = [(word, flag) for (word, flag) in cutted_words]
print(words)
```

更改使用字典

- #將結巴使用的字典更改為對繁體中文表現較好的字典
- 此字典可從結巴專案github下載



- 設定使用字典

```python
jieba.set_dictionary('./dict.txt.big')
```

```python
jieba.set_dictionary('./dict.txt.big')
```

辨識新字詞

- 啟用HMM已辨識新字詞
- 預設HMM功能即為啟用, 可以不用特地設為True

```python
input_string = "他来到了网易杭研大厦"
cutted_words = jieba.cut(input_string, HMM=True)
words = [word for word in cutted_words]
print(words)
```

```
Building prefix dict from /Users/admin/Documents/cupoy/Ch06-Jieba/dict.txt.big ...
Dumping model to file cache /var/folders/9f/mkwhwg7d4vz7rp0429_kg43c0000gn/T/jieba.u7693c7663763ede1f644a996682c676f.cache
Loading model cost 1.859 seconds.
Prefix dict has been built successfully.
['他', '来到', '了', '网易', '杭研', '大厦']
```

```python
input_string = "他来到了网易杭研大厦"
cutted_words = jieba.cut(input_string, HMM=True)
words = [word for word in cutted_words]
print(words)
```

補充資料：github

新增自定義詞庫

在既有使用的字典下新增自定義字詞

```python
jieba.load_userdict("userdict.txt")
```

```python
jieba.load_userdict("userdict.txt")
```

userdict.txt 格式如下

```
行袂開跤 2 v
袂當 4 d
袂記 4 v
袂有 4
唱著
```

動態調整字典

動態加入字典

```python
jieba.add_word('國立臺灣大學')
```

```python
jieba.add_word("國立臺灣大學")
```

動態調整詞頻

```python
jieba.suggest_freq('國立臺灣大學', True) #動態調整詞頻
```

```python
jieba.suggest_freq("國立臺灣大學", True)
```

詞性標注

進行詞性標注(PoS Tagging)

```python
import jieba
import jieba.posseg as pseg

input_str = '小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造'
cutted_words = pseg.cut(input_str)
words = [(word, flag) for (word, flag) in cutted_words]
print(words)
```

```
[('小明', 'nr'), ('碩士', 'n'), ('畢業', 'n'), ('於', 'nr'), ('國立臺灣大學', 'x'), ('，', 'x'), ('現在', 't'), ('在', 'p'), ('日本東京大學', 'nt'), ('進修', 'v'), ('深造', 'v')]
```

```python
import jieba
import jieba.posseg as pseg

input_str = "小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造"
cutted_words = pseg.cut(input_str)
words = [(word, flag) for (word, flag) in cutted_words]
print(words)
```

相關的POS對照表如下(這部分僅擷取部分)

n	名詞	名詞	取英語名詞noun的第1個字母。	人, 時, 國家
ng	名詞	名語素	名詞性語素。名詞代碼為n，語素代碼g前面置以n。	子, 身, 師, 眾
nr	名詞	人名	名詞代碼n和"人(ren)"的聲母並在一起。	連, 王, 楊

學員可以參考這個網站 查看完整對照表，也可以查看結巴官方網站

取出斷詞位置

進行詞性標注(PoS Tagging)

```python
input_str = u'小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造'
words = jieba.tokenize(input_str)
for tk in words:
    print(f'word:{tk[0]}, start:{tk[1]}, end:{tk[2]}')
```

```
word:小明, start:0, end:2
word:碩士, start:2, end:4
word:畢業, start:4, end:6
word:於, start:6, end:7
word:國立臺灣大學, start:7, end:13
word:，, start:13, end:14
word:現在, start:14, end:16
word:在, start:16, end:17
word:日本東京大學, start:17, end:23
word:進修, start:23, end:25
word:深造, start:25, end:27
```

```python
input_str = u'小明碩士畢業於國立臺灣大學，現在在日本東京大學進修深造' #在此將字串轉為unicode
words = jieba.tokenize(input_str)
for tk in words:
    print(f'word:{tk[0]}, start:{tk[1]}, end:{tk[2]}')
```

知識點回顧

這章節我們學習到了

- 使用結巴來進行各項任務(如斷詞)

詳細使用操作

請參照使用結巴進行斷詞.ipynb檔進行更詳細的使用操作

下一步：閱讀範例與完成作業