anboh AI共學社群 我的

AI共學社群 > Part1 - NLP經典機器學習馬拉松 > D11 詞性標註(Part- ... peech tagging)

D11 詞性標註(Part-of-speech tagging) 囯 範例與作業 問題討論 簡報閱讀 重要知識點 NLP自然語言學習實戰馬拉松 中文斷詞器:結巴/ Chinese word... ▶ Day11 -詞性標註(Part-of-speech tagging) **NLTK**: Natural Language Tool Kit 陪跑專家:楊哲寧 重要知識點 重要知識點 • 了解如何使用Jieba/nltk完成中英文詞性標注 Adjective Adverb Conjunction Determiner Noun Number Preposition Pronoun Verb 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:<u>結巴斷詞器與FastTag</u> jieba 斷句 複習jieba 斷句 • 中文與英文最大的差異之一,就是中文句子一般都要先經過斷句處理(英文單詞間有空格) • 先前已有課程介紹jieba斷詞,因此這裡只會快速帶過,幫助我們可以使用jieba完成PoS Tagging 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" cut_all: 調整全模式、精確模式(默認False為精確模式) HMM: 是否使用HMM算法,可以使用 HMM 模型(Hidden Markov Models)找出『未登錄詞』 : 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)))) Building prefix dict from /Users/jeff.yang/Desktop/NLP課程規劃/Chll - Part Of Speech (PoS)-2/課程練習/dict.txt ... Loading model from cache /var/folders/0f/byp88j0d48n_pnf6bpt587z80000gn/T/jieba.u942158f562cb02c8c431e7033039a2c3.cac Loading model cost 0.488 seconds. Prefix dict has been built successfully. output 精確模式: 我|愛寫|程式 output 全模式: 我愛|愛寫|程式 1 sentence_1 = "我愛寫程式" 2 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 新增單詞 In [8]: sentence_1 = "精通各種程式語言是一件相當困難的事情" In [9]: print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False)))) output 精確模式: 精通|各|種|程式|語言|是|一|件|相當|困難|的|事情 In [10]: jieba.add_word('程式語言') In [11]: print("output 精確模式: {}".format('|'.join(jieba.cut(sentence_1, cut_all=False, HMM=False)))) output 精確模式: 精通|各|種|程式語言|是|一|件|相當|困難|的|事情 1 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 我們也可以讀入整個字典,就不用一個字一個字慢慢新增 In [12]: ## 新增單詞 ,格式:每行包含一個單詞 詞類(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) In [13]: 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)))) output 精確模式: 我|在|Cupoy|平台|上|學習|自然|語言|處理 In [14]: ## 讀入字典 jieba.load_userdict('new_words.txt') In [15]: 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)))) output 精確模式: 我|在|Cupoy平台|上|學習|自然語言處理 1 ## 新增單詞 ,格式:每行包含一個單詞 詞頻(可省略) 詞性(可省略) 2 new_words = '程式語言\nCupoy平台\n自然語言處理' 3 with open('new_words.txt', 'w') as file: 4 file.write(new_words) 5 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)))) jieba.load_userdict('new_words.txt') 9 sentence_1 = "我在Cupoy平台上學習自然語言處理" 10 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 Tokenize:可以用來取出斷詞位置 In [40]: 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" In [41]: words = jieba.tokenize(sentence_2,) for tk in words: print("word: {}, start:{}, end:{}".format(tk[0],tk[1],tk[2])) word: Python, start:0, end:6 word: 是, start:6, end:7 word: 一種, start:7, end:9 word: 廣泛, start:9, end:11 word: 使用, start:11, end:13 word: 的, start:13, end:14 word: 直譯, start:14, end:16 word: 式, start:16, end:17 start:17, end:18 word: `, start:17, end:18 word: 進階, start:18, end:20 word: 程式, start:20, end:22 word: `, start:22, end:23 word: 通用型, start:23, end:26 word: 程式語言, start:26, end:30 1 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: start:{}, end:{}".format(tk[0],tk[1],tk[2])) print("word: {}, **Pos Tagging** Pos Tagging In [42]: import jieba.posseg as pseg In [43]: 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,) In [44]: ## 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) 廣泛 Vi 使用 Vt 直譯 Vt 、 x 進階 Vi 程式 N 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 Coordinating conjunction Interjection Verb, base form Proper noun, singular CD Cardinal number NNPS Proper noun, plural Verb, past tense tokenize:將句子拆成words

x w		PDT	Predeterminer	VBG	Verb, gerund or present
187	Existential there	POS	Possessive ending	participle	
	Foreign word				
N	Preposition or subordinating	PRP	Personal pronoun	VBN	Verb, past participle
conjunction		PRP\$	Possessive pronoun	VBP	Verb, non-3rd person singula
1	Adjective	RB	Adverb	present	
IR	Adjective, comparative	RBR	Adverb, comparative	VBZ	Verb, 3rd person singular
IS	Adjective, superlative	RBS	Adverb, superlative	present	
s	List item marker	RP	Particle	WDT	Wh-determiner
ΛD	Modal	SYM	Symbol	WP	Wh-pronoun
IN	Noun, singular or mass	то	to	WP\$	Possessive wh-pronoun
	reduit, singular or mass			WRB	Wh-adverb

In [4]: tokenize = nltk.word_tokenize(sentence) print('Token: ()'.format('|'.join(tokenize)))

```
Token: Wow | ... | Loved | this | place | .
         Tagging
 In [5]: tagging = nltk.pos_tag(tokenize,)
        print(tagging)
        [('Wow', 'NNP'), ('...', ':'), ('Loved', 'VBD'), ('this', 'DT'), ('place', 'NN'), ('.', '.')]
 tokenize = nltk.word_tokenize(sentence)
     print('Token: {}'.format('|'.join(tokenize)))
 3 tagging = nltk.pos_tag(tokenize,)
 4 print(tagging)
參考資料
```

網站: <u>Jieba: Github</u>

xsjy/界壩			⊙ 看 1.3千 ☆ 量	24k 学 叉子 5.9千		
① 問題 527	13. 拉取要求 47 ② 動作 🖳 專案 🗆	維基 ① 安全 ビ 見	N			
	立即加入(GitHub是超過5000萬開發人員的家園,他們共同 構建軟件]致力於託管和審查代碼,管理項目以	及共同	科個		
	註冊					
ド主・ ド2個分店 〇:	28個 横盖	轉到文件	碼 - 開於	關於		
Neutrino3316 更新READM	AE.md更新樂葉鏈接。(# 817)	67fa2e3 on 15 Feb	結巴中文分詞 口 自述文件			
extra_dict	更新到v0.33		年前 中 麻省理工學院執	er.		
■ 界端	修復python2.7中的setup.py	[python2.7中的setup.py 8				
DUE:	修復文件模式	8 f	图月前 發布 28			
.gitattributes	第一次提交			◇ v0.42.1發布 最新		
.gitignore	更新jieba3k		3年前	1月20日 + 27個版本		
□ 變更日誌	修復python2.7中的setup.py	8 f	# 27個版本 副月前			
□ 執照	添加許可證文件	1	7 年前 配套	E &		
□ 清單	在分發包中包含Changelog和README.md	7	7年前 沒有發布包			
□ 自述文件	更新README.md更新漿葉鏈接。(# 817)	7 (及角盤布包 图月前	及対策等世		
setup.py	修復python2.7中的setup.py		8月前			