自然語言機 基於jieba\_優化分詞 dict.txt

* 適用於電子商務購物網站
* 本機路徑C:\Users\User\AppData\Local\Temp\jieba.cache ，若dict.txt有新的參考時必須清除cache
* Dict.txt自訂字典格式(調整優化)

1. 多詞例如:

weather me,110,N –複合詞

weather report,110,N

1. 單詞:

防水隨身碟,300,N –品項

veloci-t,300,N --型號

gan氮化鎵pd充電器,110,N –複合詞

abyssus地獄狂蛇,110,N –複合詞

1. 詞性:N名詞

* stopword.txt停用詞的整理對於分詞結果極為重要。
* 單詞抽取:基於ms-sql資料庫自訂產生dict語法 (自訂爬取單詞或紀錄單詞table) [DBkeyword] :整理後但未洗淨過的詞字典庫; [TYPE]:詞性分類(自訂或人工審閱)

1. 品項,品牌dict匯入

SELECT keyname,300,'N',getdate() FROM [DBkeyword] rt with(nolock) where [TYPE] in (品項,品牌)

and (not PATINDEX('% %', keyname )>0)

and not ((PATINDEX('%[0-9]%', keyname )>0

and (not PATINDEX('%[一-龜]%', keyname )>0)

and (not PATINDEX('%[A-Za-z]%', keyname )>0)

))

and len(keyname)>1

and PATINDEX('%[’、。：．？！…‥·｜–‘’“”〝〞‵′＃〃,?]%', keyname)=0

and not exists (select 1 from jieba\_dict jd with(rowlock) where jd.keyword= keyname)

1. 只需要單詞;特殊條件

select keyname + ' 100 N' from [DBkeyword] (nolock) where (not PATINDEX('%.%', keyname)>0)

and [TYPE] in (品項,品牌)

and ((PATINDEX('%[A-Za-z0-9]%', keyname)>0) OR PATINDEX('%[一-龜]%', keyname)>0) --只有英文或只有中文

and (NOT ((PATINDEX('%[0-9]%', keyname)>0) AND PATINDEX('%[A-Za-z]%', keyname)=0)) --排除只有數字

order by keyname

1. 複合詞抽取

select kt.newKey\_name,300,'N',getdate() from (

select newKey\_name=REPLACE(newKey\_name,right(newKey\_name,charindex(' ',REVERSE(newKey\_name))),'')

from (

SELECT newKey\_name=REPLACE(Keyname,right(Keyname,charindex(' ',REVERSE(Key\_name))),'')

FROM [DBkeyword] kt with(nolock) where [TYPE]= 複合詞

and PATINDEX('% %', Keyname )>0 and PATINDEX('%[一-龜]%', Keyname )=0

and PATINDEX('[A-Za-z]% [0-9]%', Keyname )>0

group by REPLACE(Keyname,right(Keyname,charindex(' ',REVERSE(Keyname))),'')

)kt) kt

where len(kt.newKey\_name)>=6 and

not kt.newKey\_name like '%-%'

and (PATINDEX('[0-9]%', kt.newKey\_name )=0)

and (PATINDEX('[A-Za-z][A-Za-z][0-9]%', kt.newKey\_name )=0)

and (PATINDEX('[A-Za-z][0-9]%', kt.newKey\_name )=0)

and (PATINDEX('[A-Za-z][0-9][0-9]%', kt.newKey\_name )=0)

and (PATINDEX('[A-Za-z][0-9][0-9][0-9]%', kt.newKey\_name )=0)

and not exists (select 1 from new\_jieba\_dict jd with(nolock) where jd.keyword=kt.newKey\_name) group by kt.newKey\_name;

* 優化過的程序段落

Jieba\\_\_init\_\_.py : 調整取以 “,” 區隔

☆ jieba\posseg\\_\_init\_\_.py : 調整許多正規表示法及自訂函數字串單詞、複合詞、英數單詞、品項規格、型號的抽取 (重要調整)

Ex:

#ipjone 13 or pjone 7 or htc 10pro

re\_ban1\_1 =re.findall(r'[a-zA-Z]+\s+[0-9]+\s+[a-zA-Z]+|[a-zA-Z]+\s+[0-9]+[a-zA-Z]\*',sentence)

#取出型號規則(不含-)ex: def789ghi or def456

#3m or 4PQWER

#QA82Q900RBWXZW

re\_ban1=re.findall(r'[A-Za-z]+[0-9]+?[A-Za-z0-9]\*|[0-9]+[A-Za-z]+?[A-Za-z0-9]\*|[A-Za-z]+[0-9]+[A-Za-z]+|[A-Za-z]+[0-9]\*|[a-zA-Z]+',sentence)

#取出ex: abc-123

re\_ban0=re.findall(r'[A-Za-z0-9]+-[A-Za-z0-9]+-?[A-Za-z0-9]\*',s)

* 一些特殊的表示式調整

re\_han\_detail = re.compile("([\u4E00-\u9FD5]+)")#中文

re\_skip\_detail = re.compile("([\.0-9]+|[a-zA-Z0-9]+)")

re\_han\_internal = re.compile("([\u4E00-\u9FD5a-zA-Z0-9+#&\.\_]+)")#中英文 和 特殊符號

re\_skip\_internal = re.compile("(\r\n|\s)")

re\_eng = re.compile("[a-zA-Z0-9]+")

re\_num = re.compile("[\.0-9]+")

re\_eng\_nc = re.compile("[a-zA-Z]+")#只有英文的單詞

re\_eng1 = re.compile('^[a-zA-Z0-9]$', re.U)

re\_eng\_num =re.compile("([A-Za-z]+\d+|\d+[A-Za-z]+)[A-Za-z0-9]\*$")#英數混和

#只取英數型號

def filter\_pno(sentence):

#去掉多餘的空行

blank\_line=re.compile('\n+')

re\_br=re.compile('<br\s\*?/?>')#處理換行

re\_h=re.compile('</?\w+[^>]\*>')#HTML標籤

re\_comment=re.compile('<!--[^>]\*-->')#HTML注釋

s=re\_br.sub('',sentence)#去掉<br

s=re\_h.sub('',s) #去掉HTML標籤

s=re\_comment.sub('',s)#去掉HTML注釋

s=blank\_line.sub('',s)

#取出ex: abc-123

re\_ban0=re.findall(r'[A-Za-z0-9]+-[A-Za-z0-9]+-?[A-Za-z0-9]\*',s)

#取出型號規則(不含-)ex: def789ghi or def456

#3m or 4PQWER

#QA82Q900RBWXZW

re\_ban1=re.findall(r'[A-Za-z]+[0-9]+?[A-Za-z0-9]\*|[0-9]+[A-Za-z]+?[A-Za-z0-9]\*|[A-Za-z]+[0-9]+[A-Za-z]+|[A-Za-z]+[0-9]\*|[a-zA-Z]+',s)

items1 = []

for keyword\_0 in re\_ban0:

if (len(keyword\_0) >=2):

items1.append(keyword\_0.strip())

for keyword\_1 in re\_ban1:

finds =""

for i,x in enumerate(items1):

if (keyword\_1.find(x)!=-1) or ((x.find(keyword\_1)!=-1) and (len(x)>(len(keyword\_1)))):

finds = x

break;

if (finds=="") and (len(keyword\_1) >=2) :

items1.append(keyword\_1.strip())

if (len(items1)>0) :

list2 = list(set(items1)) #消除重複的元素

return list2

else: return ""

jieba\analyse\\_\_init\_\_.py :停用詞函數優化

jieba\analyse\analyzer.py、tfidf.py : 停用詞函數優化STOP\_WORDS = frozenset(( ….. ))

jieba\analyse\textrank.py : 詞性調整 增加效率(只取需要的) textrank()