GitHub Gist

luw2007 / 詞性標記.md
Last active 18 days ago

詞性標記： 包含ICTPOS3.0詞性標記集、ICTCLAS 漢語詞性標註集、jieba 字典中出現的詞性、simhash 中可以忽略的部分詞性

詞性標記.md

詞的分類

- 實詞：名詞、動詞、形容詞、狀態詞、區別詞、數詞、量詞、代詞
- 虛詞：副詞、介詞、連詞、助詞、擬聲詞、嘆詞。

ICTPOS3.0詞性標記集

```
n 名词
    nr 人名
        nr1 汉语姓氏
        nr2 汉语名字
        nrj 日语人名
        nrf 音译人名
    ns 地名
      nsf 音译地名
    nt 机构团体名
    nz 其它专名
    nl 名词性惯用语
    ng 名词性语素

t 时间词
   tg 时间词性语素

s 处所词

f 方位词

v 动词
    vd 副动词
    vn 名动词
    vshi 动词“是”
    vyou 动词“有”
    vf 趋向动词
    vx 形式动词
    vi 不及物动词（内动词）
    vl 动词性惯用语
    vg 动词性语素
a 形容词
    ad 副形词
    an 名形词
    ag 形容词性语素
    al 形容词性惯用语
b 区别词
    bl 区别词性惯用语
z 状态词
r 代词
    rr 人称代词
    rz 指示代词
        rzt 时间指示代词
        rzs 处所指示代词
        rzv 谓词性指示代词
    ry 疑问代词
        ryt 时间疑问代词
        rys 处所疑问代词
        ryv 谓词性疑问代词
    rg 代词性语素
m 数词
    mq 数量词
```

```
q 量词
    qv 动量词
    qt 时量词
```

虛詞

```
d 副词
p 介词
    pba 介词“把”
    pbei 介词“被”
c 连词
    cc 并列连词
u 助词
    uzhe 着
    ule 了 喽
    uguo 过
    ude1 的 底
    ude2 地
    ude3 得
    usuo 所
    udeng 等 等等 云云
    uyy 一样 一般 似的 般
    udh 的话
    uls 来讲 来说 而言 说来

    uzhi 之
    ulian 连 （“连小学生都会”）

e 叹词
y 语气词(delete yg)
o 拟声词
h 前缀
k 后缀
x 字符串
    xx 非语素字
    xu 网址URL
w 标点符号
    wkz 左括号，全角：（ 〔  [  ｛  《 【  〖 〈   半角：( [ { <
    wky 右括号，全角：） 〕  ] ｝ 》  】 〗 〉 半角： ) ] { >
    wyz 左引号，全角：“ ‘ 『
    wyy 右引号，全角：” ’ 』
    wj 句号，全角：。
    ww 问号，全角：？ 半角：?
    wt 叹号，全角：！ 半角：!
    wd 逗号，全角：， 半角：,
    wf 分号，全角：； 半角： ;
    wn 顿号，全角：、
    wm 冒号，全角：： 半角： :
    ws 省略号，全角：…… …
    wp 破折号，全角：——   －－   ——－   半角：---  ----
    wb 百分号千分号，全角：％ ‰   半角：%
    wh 单位符号，全角：￥ ＄ ￡  ° ℃  半角：$
```

ICTCLAS 漢語詞性標註集

代碼	名稱	幫助記憶的詮釋
Ag	形語素	形容詞性語素。形容詞代碼為a，語素代碼 g 前面置以A。
a	形容詞	取英語形容詞adjective的第1個字母。
ad	副形詞	直接作狀語的形容詞。形容詞代碼a和副詞代碼d並在一起。
an	名形詞	具有名詞功能的形容詞。形容詞代碼a和名詞代碼n並在一起。
b	區別詞	
c	連詞	
Dg	副語素	

代碼	名稱	建議更好的譯法
d	副詞	取adverb的第2個字母，因其第1個字母已用於形容詞。
e	嘆詞	取英語嘆詞exclamation的第1個字母。
f	方位詞	取漢字“方” 的聲母。
g	語素	絕大多數語素都能作為合成詞的“詞根”，取漢字“根”的聲母。
h	前接成分	取英語head的第1個字母。
i	成語	取英語成語idiom的第1個字母。
j	簡稱略語	取漢字“簡”的聲母。
k	後接成分	
l	習用語	習用語尚未成為成語，有點“臨時性”，取“臨”的聲母。
m	數詞	取英語numeral的第3個字母，n，u已有他用。
Ng	名語素	名詞性語素。名詞代碼為n，語素代碼 g 前面置以N。
n	名詞	取英語名詞noun的第1個字母。
nr	人名	名詞代碼n和“人(ren)”的聲母並在一起。
ns	地名	名詞代碼n和處所詞代碼s並在一起。
nt	機構團體	“團”的聲母為t，名詞代碼n和t並在一起。
nz	其他專名	“專”的聲母的第1個字母為z，名詞代碼n和z並在一起。
o	擬聲詞	取英語擬聲詞onomatopoeia的第1個字母。
p	介詞	取英語介詞prepositional的第1個字母。
q	量詞	取英語quantity的第1個字母。
r	代詞	取英語代詞pronoun的第2個字母,因p已用於介詞。
s	處所詞	取英語space的第1個字母。
Tg	時語素	時間詞性語素。時間詞代碼為t,在語素的代碼g前面置以T。
t	時間詞	取英語time的第1個字母。
u	助詞	取英語助詞auxiliary 的第2個字母,因a已用於形容詞。
Vg	動語素	動詞性語素。動詞代碼為v。在語素的代碼g前面置以V。
v	動詞	取英語動詞verb的第一個字母。
vd	副動詞	直接作狀語的動詞。動詞和副詞的代碼並在一起。
vn	名動詞	指具有名詞功能的動詞。動詞和名詞的代碼並在一起。
w	標點符號	
x	非語素字	非語素字只是一個符號，字母x通常用於代表未知數、符號。
y	語氣詞	取漢字“語”的聲母。
z	狀態詞	取漢字“狀”的聲母的前一個字母。

jieba 字典中出現的類型

```
a
    ad
    ag
    an
b
c
d
    df
```

```
    dg
e
f
g
h
i
j
k
l
m
    mg
    mq
n
    ng
    nr
    nrfg
    nrt
    ns
    nt
    nz
o
p
q
r
    rg
    rr
    rz
s
t
    tg
u
    ud
    ug
    uj
    ul
    uv
    uz
v
    vd
    vg
    vi
    vn
    vq
x
y
z
    zg
```

simhash 過濾的掉的詞彙

```
c
e
h
k
o
p
u
    ud
    ug
    uj
    ul
    uv
    uz
x
y
```