清华大学中文分词词性标注**THULAC**

**一、下载地址：**<http://thulac.thunlp.org/>

**二、简介**

THULAC（THU Lexical Analyzer for Chinese）由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。THULAC集成了目前世界上规模最大的人工分词和词性标注中文语料库（约含5800万字），模型标注能力强大。该工具包在标准数据集Chinese Treebank(CTB5)上分词的F1值可达97.3%，词性标注的F1值可达到92.9%。同时进行分词和词性标注时每秒可处理约15万字，速度较快。

**三、使用教程**

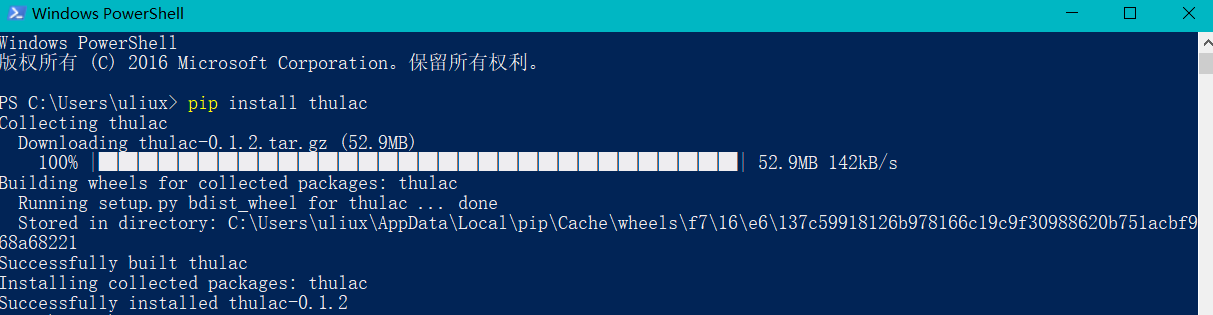
**1、编译和安装（基于python版本，兼容2.x和3.x）**

**1.源代码下载**

将thulac文件放到目录下，通过import thulac来引用，thulac需要模型的支持，需要将下载的模型放到thulac目录下。

**2.pip下载**

进入powershell命令行模式，输入pip install thulac，然后回车，通过import thulac来引用。如下图：



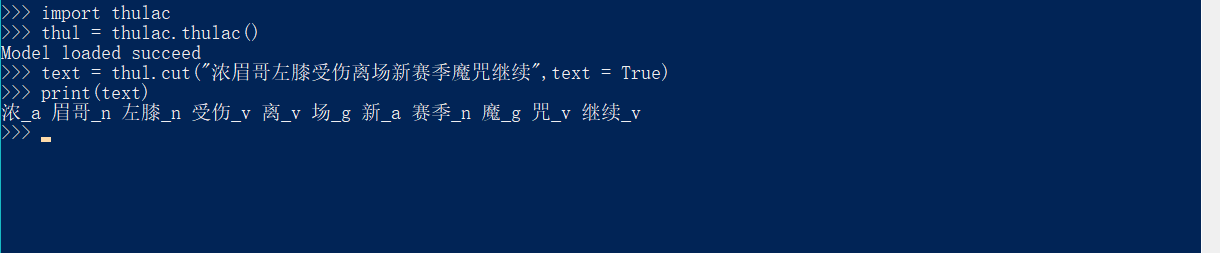
**2、使用方式**

**⒈分词和词性标注程序**

**1.1命令格式（Python 3.x版）**

通过python程序导入 thulac包，新建thulac.thulac(args)类，其中args为程序的参数。之后可以通过调用thilac.cut()进行单句分词。注意：thulac只支持GBK和ASII格式文档。

代码示例1:



代码示例2：



**1.2 python版本接口参数**

◎thulac(user\_dict=None,model\_path=None,T2S=False, seg+only=False)初始化程序，进行自定义设置。其中：

user\_dict：设置用户字典，用户字典中的次会被打上uw标签。词典中每一个词一行，UTF-8编码。

T2S：默认False，是否将句子从繁体转化为简体

Seg\_only：默认False,是否只进行分词，不进行词性标注

Filt：默认False,是否使用过滤器去除一些没有意义的词语，例如“可以”。

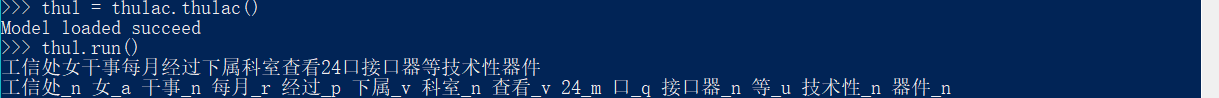
model\_path：设置模型文件所在文件夹，默认为models/

◎cut(文本，text=False)：对一句话进行分词

text：默认为False,是否返回文本，不返回文本则返回一个二维数组（[[word,tag]…]）,tag\_only模式下tag为空字符。

◎cut\_f：(输入文件，输出文件)对文件进行分词

◎run()：命令行交互式分词（屏幕输入、屏幕输出）,如图：



**1.3分词和词性标注模型的使用**

THULAC需要分词和词性标注模型的支持，用户可以下载THULAC模型Models\_v1.zip，并放到THULAC的根目录下即可，或者使用参数——model\_dir dir指定模型的位置。

**2.词性标记集**

**2.1通用标记集：**

n/名词 np/人名 ns/地名 ni/机构名 nz/其它专名m/数词 q/量词

mq/数量词 t/时间词 f/方位词 s/处所词v/动词 a/形容词 d/副词

h/前接成分 k/后接成分 i/习语 j/简称 r/代词 c/连词 p/介词

u/助词 y/语气助词 e/叹词 o/拟声词 g/语素 w/标点 x/其它

**2.2特殊标记集(适用于lite\_v1\_2版)**

需要下载使用：

vm/能愿动词 vd/趋向动词

**四、相关论文**

Zhongguo Li, Maosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguistics, vol. 35, no. 4, pp. 505-512, 2009.