附录：中文文本分析

2023 年 9 月 18 日

中文文本分析——Jieba

Jieba，是一个流行的 Python 中文分词组件。

1. 安装 Jieba

如果使用 pip 安装工具，在终端 Terminal 里的命令行模式下，输入以下命令：

```
pip install jieba
```

或使用 ipython 的魔术命令 (magic commands)直接运行以下代码，

[ ]:
```
%pip install jieba
```

2. 分词

[16]:
```
import jieba
text="《再别康桥》轻轻的我走了，正如我轻轻的来；我轻轻的招手，作别西天的云彩。那河畔的金柳，是夕阳中的新娘；波光里的艳影，在我的心头荡漾。软泥上的青荇 ，油油的在水底招摇 ；在康河的柔波里，我甘心做一条水草！那榆荫下的一潭，不是清泉，是天上虹；揉碎在浮藻间，沉淀着彩虹似的梦。寻梦？撑一支长篙 ，向青草更青处漫溯 ；满载一船星辉，在星辉斑斓里放歌。但我不能放歌，悄悄是别离的笙箫；夏虫也为我沉默，沉默是今晚的康桥！悄悄的我走了，正如我悄悄的来；我挥一挥衣袖，不带走一片云彩。"
seg_list = jieba.cut(text, cut_all=True)
print("/ ".join(seg_list))  # 全模式
```

《/ 再别/ 再别康桥/ 别康桥/ 康桥/ 》/ 轻轻/ 的/ 我/ 走/ 了/ ，/ 正如/ 我/ 轻轻/ 的/
↪来/ ；/ 我/ 轻轻/ 的/ 招手/ ，/

作别/ 西天/ 的/ 云彩/ 。/ 那/ 河畔/ 的/ 金/ 柳/ ，/ 是/ 夕阳/ 中/ 的/ 新娘/ ；/ 波光/ 里/ 的/ 艳/ 影/ ，/ 在/
我/ 的/ 心头/ 荡漾/ 。/ 软泥/ 上/ 的/ 青/ 荇/ ，/ 油油/ 油油的/ 在/ 水底/ 招摇/ ；/ 在/ 康/ 河/ 的/ 柔波/ 波里/
，/ 我/ 甘心/ 做/ 一条/ 水草/ ！/ 那/ 榆/ 荫/ 下/ 的/ 一/ 潭/ ，/ 不是/ 清泉/ ，/是/ 天上/ 虹/ ；/ 揉碎/ 在/
浮/ 藻/ 间/ ，/ 沉淀/ 着/ 彩虹/ 似的/ 梦/ 。/ 寻梦/ ？/ 撑/ 一支/ 长/ 篙/ ，/ 向/ 青草/ 更/ 青/ 处/ 漫/ 溯/
；/ 满载/ 一/ 船/ 星/ 辉/ ，/ 在/ 星/ 辉/ 斑斓/ 里/ 放歌/ 。/ 但/ 我/ 不能/ 放歌/ ，/ 悄悄/ 是/ 别离/ 的/ 笙箫/
；/ 夏/ 虫/ 也/ 为/ 我/ 沉默/ ，/ 沉默/ 是/ 今晚/ 的/ 康桥/ ！/ 悄悄/ 悄悄的/ 我/走/ 了/ ，/ 正如/ 我/ 悄悄/
悄悄的/ 来/ ；/ 我/ 挥/ 一/ 挥/ 衣袖/ ，/ 不/ 带走/ 一片/ 云彩/ 。

3. 词性标注

[17]:
```
import jieba.posseg as pseg
words = pseg.cut("轻轻的我走了，正如我轻轻的来") #jieba 默认模式
for word, flag in words:
    print('%s %s' % (word, flag))
```

```
轻轻 d
的 uj
我 r
走 v
了 ul
， x
正如 v
我 r
轻轻 d
的 uj
来 v
```

词性和专名类别标签集合如下表:

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	t	时间

标签	含义	标签	含义	标签	含义	标签	含义
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

参考

Jieba：https://github.com/fxsjy/jieba