

## 《知识图谱》实验报告

学 号： 1005183121姓 名： 周子杰日 期： 2021. 6. 23

得 分： \_\_\_\_\_

### 一、 实验内容

- ✧ 命名实体识别
- ✧ 词性标注
- ✧ 中文分词

### 二、 实现方法

#### 1、命名实体识别

##### 基础概念：

命名实体识别是指从非结构化的文本中抽取出已命名的实体，包括数量、时间等简单实体和人名、地名、机构名等相对困难的实体。其被广泛应用在知识图谱、机器翻译等领域。

实现的方法主要有四个：

- ✧ 基于规则和词典的方法
- ✧ 基于统计的方法
- ✧ 混合方法
- ✧ 基于神经网络的方法

##### 代码实现：

在代码的实现上，我们引用 jieba 库，在这个库的基础上实现全模式、精确模式、搜索引擎模式和关键词提取，这里主要用到的是 jieba 中的 cut 函数。

同时，也可以自己加入一些新词，采用 jieba 中的 add\_word 函数即可实现该功能。jieba 库中具体的实现代码在这里就不再展示了。

```
1. jieba.add_word('石墨烯')
2. terms = jieba.cut('python 的正则表达式是好用的')
```

## 2、词性标注

### 基础概念：

词性标注是为词串中的词赋予词性标记。其难点主要有兼类现象和两类约束（局部约束、上下文约束）。

上课时介绍的方法有三个：

- ✧ 基于 HMM 的词性标注
- ✧ 基于转换的词性标注
- ✧ 基于分类思想的词性标注

### 代码实现：

这里的实现方式还是采用 jieba，先用 cut 函数进行实体识别操作，然后对识别出来的每一个单词进行词性的输出。和上面一样，jieba 库中的代码就不再展示了。

```
1. text = "去北京大学玩"
2. seg = psg.cut(text)
3. for ele in seg: # 将词性标注结果打印出来
4.     print(ele)
```

## 3、中文分词

### 基础概念：

分词就是将一句话切成一个个单词的过程，其目的是更加有效、准确的关键词索引。其应用领域非常广泛，如汉字处理、信息检索等。中文的词性标注相比印欧语缺少词形态变化，很难从词的形态变化上来判别。

中文分词的常用方法有：

- ✧ 基于词典的分词法：最大匹配法、最少分词法（最短路径法）
- ✧ 基于统计的分词法：生成式统计分词、判别式统计分词
- ✧ 基于理解的分词法

### 代码实现：

实现的方法还是利用 jieba 库，用 cut 函数来进行分词，对于特殊的词汇，调用 add\_word 来添加自定义词典或者调整词典，用 analyse 中的方法来进行关

键字提取，用 tokenize 来返回词语在原文的起止位置等等。限于篇幅，jieba 库中的具体实现方法同上不展示了。

### 三、 结果分析

#### 1、命名实体识别

实验结果：

```
【全模式】：他/ 来到/ 上海/ 上海交通大学/ 交通/ 大学/ 吃/ 蜜/ 雪/ 冰城
【精确模式】：他/ 来到/ 上海交通大学
【搜索引擎模式】：他/ 毕业/ 于/ 上海/ 交通/ 大学/ 上海交通大学/ 机电/ 系/ ， / 后来/ 在/ 一部/ 上海 / 电器/ 科学/ 研究/ 研究所/ 工作
欧亚 0.7300142700289363
吉林 0.659038184373617
置业 0.4887134522112766
万元 0.3392722481859574
增资 0.33582401985234045
```

结果说明：

这是全模式、精确模式、搜索引擎模式和关键词提取的结果，可以看到识别的准确率还是相当好的。

#### 2、词性标注

实验结果：

```
去/v
北京大学/nt
玩/v
```

结果说明：

其中，v 代表动词，nt 代表机构团体，可以看到结果是正确无误的。

#### 3、中文分词

实验结果：

```
=====
1. 分词

Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\24763\AppData\Local\Temp\jieba.cache
Loading model cost 0.495 seconds.
Prefix dict has been built successfully.
全模式（切分出字典里所有的词）：六月/ 25/ 号/ 傍晚/ 晚会/ 在/ 奥林匹/ 奥林
默认模式（同一段字只选择词频最高的）：六月/ 25/ 号/ 傍晚/ 会/ 在/ 奥林匹克
六月，25，号，傍晚，会，在，奥林匹克公园，的，鸟巢，放，烟花，，，大家，一
昨天，我们，听，了，，，中国，科学，学院，科学院，中国科学院，周，院士，的
=====
2. 添加自定义词典/调整词典

你/爱/我/呀/我爱你/，/蜜/雪/冰城/甜蜜蜜
Before: None, After: 1
你/爱/我/呀/我爱你/，/蜜雪/冰城/甜蜜蜜
[ /台/中/ ] /正确/的话/应该/不会/被/切开
Before: None, After: 69
[ /台/中/ ] /正确/的话/应该/不会/被/切开
夫/鸚/鵒/发/于/南海/，/而/飞于/北海/；/非/梧桐/不止/，/非/练/实不食/，/非/體
夫/鸚/鵒/发/于/南海/，/而/飞于/北海/；/非/梧桐/不止/，/非/练/实不食/，/非/體
夫/鸚/鵒/发/于/南海/，/而/飞于/北海/；/非/梧桐/不止/，/非/练/实不食/，/非/體
夫/鸚/鵒/发/于/南海/，/而/飞于/北海/；/非/梧桐/不止/，/非/练/实不食/，/非/體
=====
```

3. 关键词提取

TF-IDF 算法

欧亚 0.7300142700289363  
吉林 0.659038184373617  
置业 0.4887134522112766  
万元 0.3392722481859574  
增资 0.33582401985234045  
4.3 0.25435675538085106  
7000 0.25435675538085106  
2013 0.25435675538085106  
139.13 0.25435675538085106  
实现 0.19900979900382978  
综合体 0.19480309624702127  
经营范围 0.19389757253595744  
亿元 0.1914421623587234  
在建 0.17541884768425534  
全资 0.17180164988510638  
注册资本 0.1712441526  
百货 0.16734460041382979  
零售 0.1475057117057447  
子公司 0.14596045237787234  
营业 0.13920178509021275

TextRank 算法

吉林 1.0  
欧亚 0.9966893354178172  
置业 0.6434360313092776  
实现 0.5898606692859626  
收入 0.43677859947991454  
增资 0.4099900531283276  
子公司 0.35678295947672795  
城市 0.34971383667403655  
商业 0.34817220716026936  
业务 0.3092230992619838  
在建 0.3077929164033088  
营业 0.3035777049319588  
全资 0.303540981053475  
综合体 0.29580869172394825  
注册资本 0.29000519464085045  
有限公司 0.2807830798576574  
零售 0.27883620861218145  
百货 0.2781657628445476  
开发 0.2693488779295851  
经营范围 0.2642762173558316

启用停用词之后的结果:

置业 0.5889623654853846  
万元 0.4088665555061538  
增资 0.40471099828358975  
4.3 0.30653250007435895  
7000 0.30653250007435895  
2013 0.30653250007435895  
139.13 0.30653250007435895  
实现 0.23983232187641024  
综合体 0.23476270573358973  
经营范围 0.23367143356897435  
亿元 0.23071234950923078  
在建 0.21140220105538463  
全资 0.20704301396410257  
注册资本 0.20637115826153846  
百货 0.20167169793461537  
零售 0.17776329359410256  
子公司 0.17590105799384614  
营业 0.16775599741641026  
净利润 0.1535356008023077  
商业 0.1479948079448718  
:10

搜索模式

word 永和

nd:2

start: 0

end:2

:4

word 服装

nd:4

start: 2

end:6

:10

word 饰品

nd:6

start: 4

end

word 有限

nd:8

start: 6

end

:2

word 公司

start: 8

end:10

word 有限公司

start: 6

end:10

结果说明:

可以看到，各个功能的结果正确。在加入了自定义词典后，对中文的识别非常准确。

四、 结论与展望

本次实验让我受益匪浅，学习到了很多关于知识图谱的知识，包括实体识别、词性标注和中文分词等等。学姐学长为我们讲述了这些功能的概念、原理和实现方法，有一点类似之前学习的编译原理这门课，所以理解起来也相对容易。

当然，由于这里的语言是 python，所以我也更加理解了 python 的一些知识。包括如何去运用 jieba 库。实际上在寒假参加数学建模美赛的时候我已经很浅地接触到了 jieba 这个组件，非常幸运能够在这几天去学习关于这个组件的一些原理相关的知识。

在这些知识之外我们还听了周成虎院士的讲话，其关于 gis 的发展等的叙述也让我印象深刻，能够听到这样的一场讲座是我的荣幸，非常感谢学校提供给我们这样的机会。