

26、文本数据处理

1、文本数据处理的概念

1) 文本数据处理是大数据处理的重要分支之一，目的是从大规模的文本数据中提取出符合需要的、感兴趣的和隐藏的信息。目前，文本数据处理主要应用在搜索引擎、情报分析、自动摘要、自动校对、论文查重、文本分类、垃圾邮件过滤、机器翻译、自动应答等方面。

2) 文本内容是非结构化的数据，要从大量的文本中提取出有用的信息，需要将文本从无结构的原始状态转化为结构化的、便于计算机处理的数据。典型的文本处理过程主要包括：分词、特征提取、数据分析、结果呈现等。

3) 因为英文词语和词语之间有明显的空格，而中文词和词之间是紧密相连的，这就使得中文分词方法相当复杂，目前常用的中文词法分类有基于词典的分词方法、基于统计的分词方法、基于规则的分词方法。常见的分词系统有jieba、IKAnalyzer、NLPIR、语言云、BosonNLP等，其中jieba是基于python的一个开源项目。

2、文本数据分析的应用

1) 标签云。用词频表现文本特征，以文字大小的形式代表词语的重要性。

2) 文本情感分类。对文本的主观性、观点、情绪、极性进行挖掘和分析，对文本的情感倾向作出分类判断。

3、制作标签云

在网上搜索《“十四五”时期经济社会发展指导方针》，并将文本内容保存到“十四五时期经济社会发展指导方针.txt”文件中。请回答以下问题：

1) 文件“十四五时期经济社会发展指导方针.txt”中的数据是否是结构化数据：____（填：是/否）。

2) 以下Python程序是利用jieba将该文本内容进行分词，并得出出现次数最高的词。请完善程序代码：

```
import jieba
with open("十四五时期经济社会发展指导方针.txt",'r') as f:
    s=f.read()
text_list=jieba.cut(s)
text_count={}
for t in text_list:
    if len(t)==1:
        continue #当该词为单字时结束本次循环，继续下一次循环
    text_count[t]=text_count.get(t,0)+1 #get(t,0)表示获取键名为t的值，若没有该键则得到0
words=list(count.items())
words=sorted(words,key=lambda x:x[1],reverse=True)
```

3) 生成标签云

将文本内容分词后，就可以使用wordcloud生成标签云了。过程包括：配置参数、生成标签云、显示图片。

1) 导入库与配置参数:

```
from wordcloud import WordCloud, STOPWORDS #导入wordcloud模块中相关子包
import numpy as np
from PIL import Image
wc=WordCloud(
    background_color='white', #背景色参数
    max_words=100, #显示词数量
    mask=np.array(Image.open("mask.png")), #读取图片文件作为标签云轮廓
    stopwords=STOPWORDS, #使用默认的停用词,使用停用词的目的是避免出现常见的介词等
    font_path='simhei.ttf', #设置字体
    max_font_size=200 #设置最大字号
)
```

2) 生成标签云:

```
wc.fit_words(text_count) #使用词频字典生成标签云
```

```
<wordcloud.wordcloud.WordCloud at 0x7fcdf37073c8>
```

3) 显示图片:

```
import matplotlib.pyplot as plt
plt.imshow(wc)
plt.axis('off')
plt.show()
```

