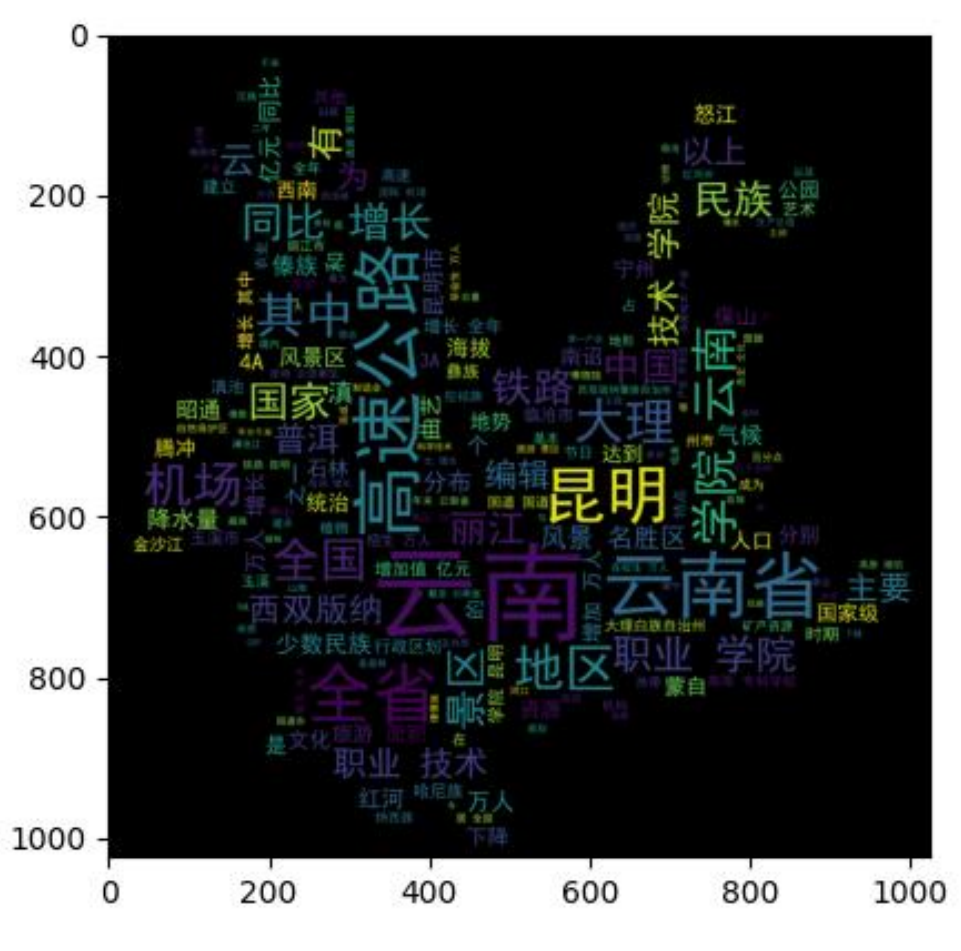
基于PYTHON可视化分析《西游记》文本

PYTHON分析《西游记》中的人物出场情况

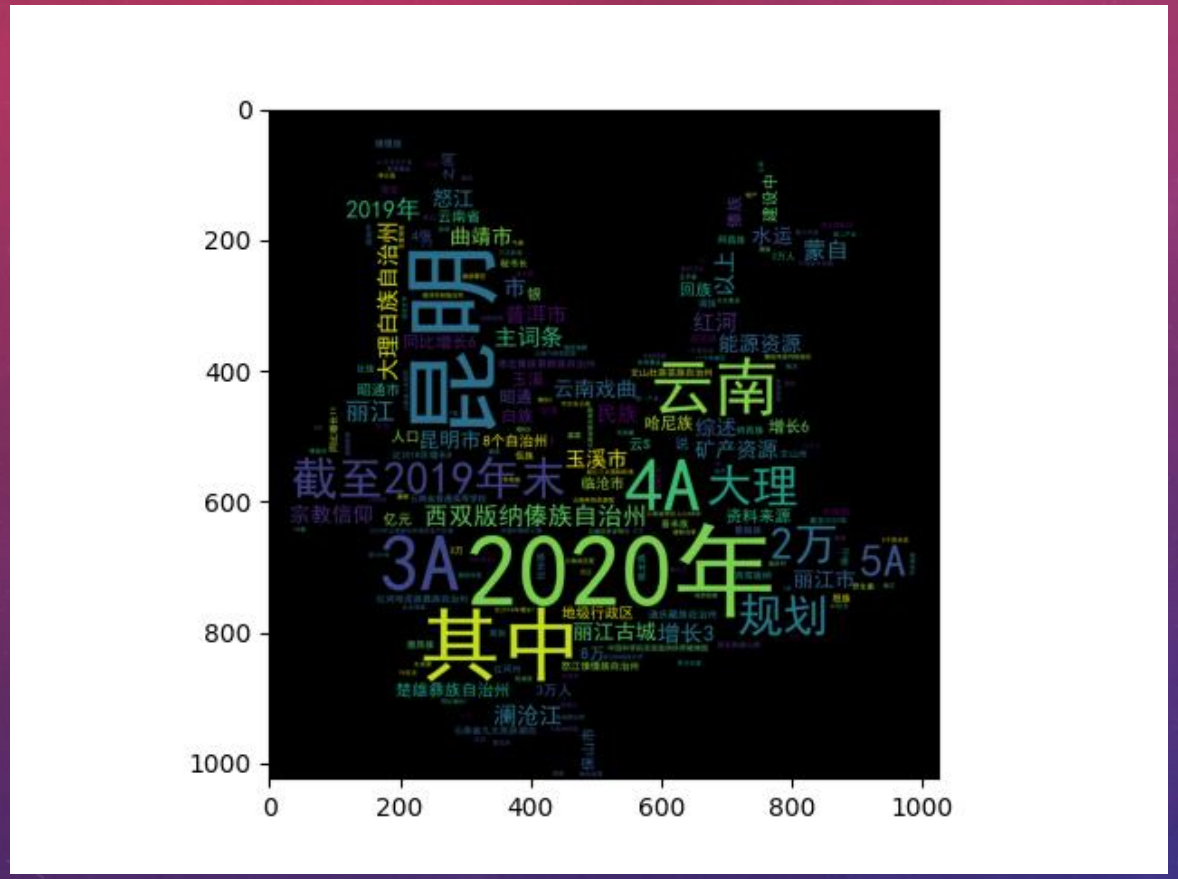
昆明市第一中学西山学校 齐洪

知识回顾——对云南词条的分析

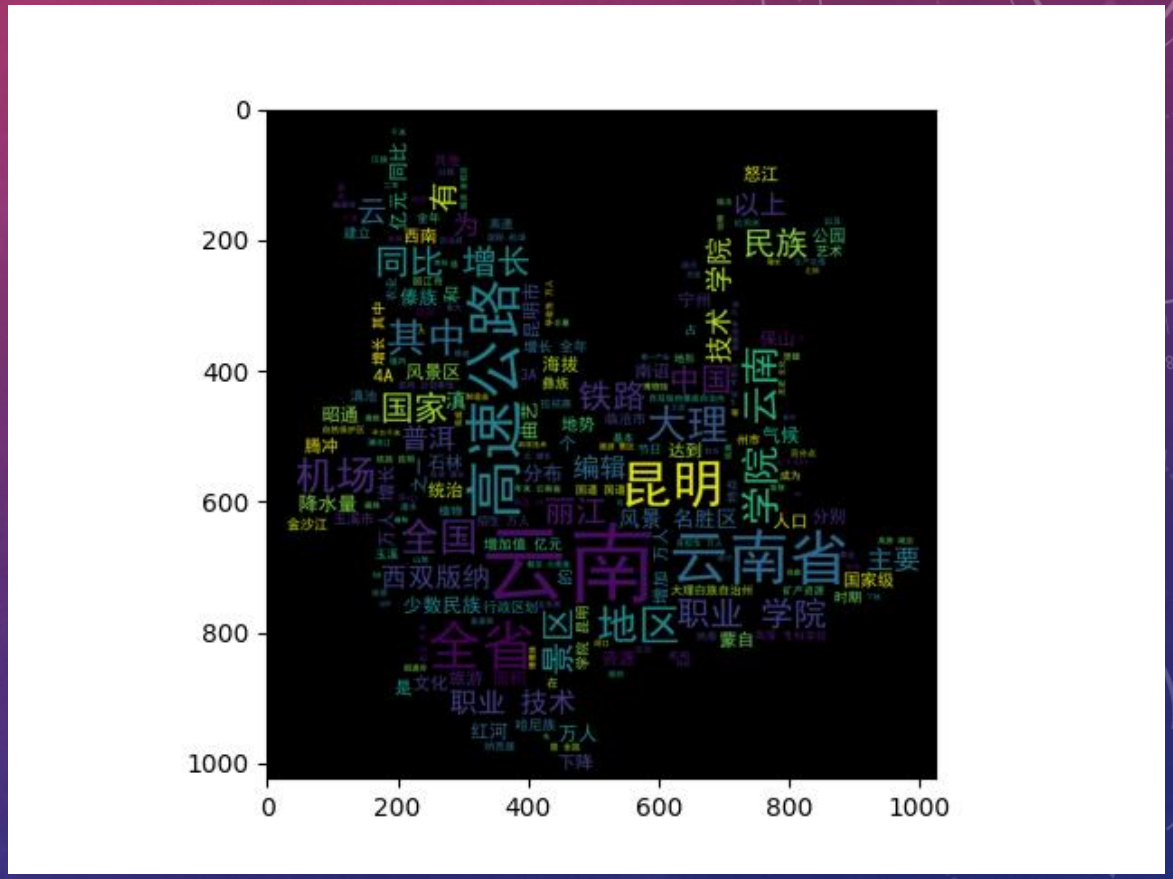



```
#导入外部库，matplotlib.pyplot是绘图库
#wordcloud是词云制作库
import numpy as np
from PIL import Image
import matplotlib.pyplot  as plt
import wordcloud as wc
import jieba

#先将数据文件打开并读取
f1=open("yunnan.txt",encoding="utf-8")
fs1=f1.read()
f1.close()

#形成词云
mask = np.array(Image.open("yunnan.jpg"))
a=wc.WordCloud(font_path="simhei.ttf",mask=mask).generate(fs1)

#显示并保存词云
plt.imshow(a)
plt.savefig(fname="云南词条分析1.png")
plt.show()
```

中文词云制作时，为什么需要**先分词**再制作词云？



jieba分词之前



jieba分词之后

中文分词的主要方法

分词工具都不能做到百分之百准确，选择不同分词工具或外部库需要理解它的缺陷与特点。

- **基于字符串匹配的分词方法：**按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。
- **基于理解的分词方法：**在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。
- **基于统计的分词方法：**词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。
- **基于统计机器学习的方法：**首先给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。

Python处理数据及词云可视化的过程

数据收集 → 数据整理 → 数据分析 → 数据呈现

导入外部处理库 → 文本导入与读取 → 数据整理与分词 → 特征提取形成词云 → 显示词云并保存

- 如果我们要处理的文本不是几千字，还是几十万，几百万，甚至几亿的数据量，会有怎样的要求？
- 从循环结构去思考，前面代码中删除单个字符的代码有什么问题？

```
#循环中将长度小于2的字符删除
for w in fs:
    if len(w) == 1:
        fs.remove(w)
```

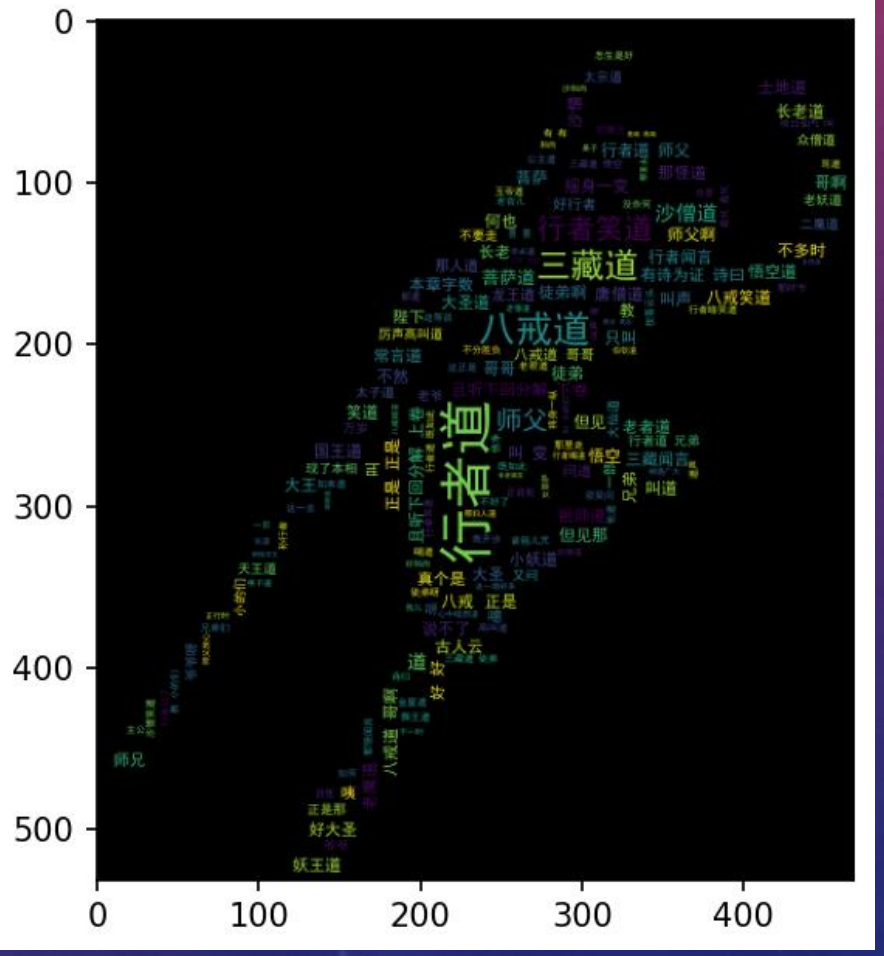
大数据

大数据的概念	大数据是指无法在可承受的时间范围内用常规软件工具进行高效捕捉、管理和处理的信息集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

- 大数据（big data），数据量大，**无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。**
  - 体量大
  - 数据类型繁多
  - 变化数据快
  - 价值密度低但收益广

学习活动1：对西游记文本分词并可视化

```
#导入数据分析的外部库
import matplotlib.pyplot  as plt
import wordcloud as wc
import numpy as np
from PIL import Image

#先将数据文件打开并读取
f1=open("西游记.txt", encoding="ANSI")
xiyou=f1.read()
f1.close()

#形成词云
bg=np.array(Image.open("1.jpg"))
ciyun=wc.WordCloud(font_path="simhei.ttf",mask=bg)
ciyun.generate(xiyou)

#显示并保存词云
plt.imshow(ciyun)
plt.savefig(fname="西游记文本分析1.png")
plt.show()
```



（1）通过jieba分词库对中文进行分词

words = jieba.lcut(xiyou)

（2）通过join连接空格和列表元素

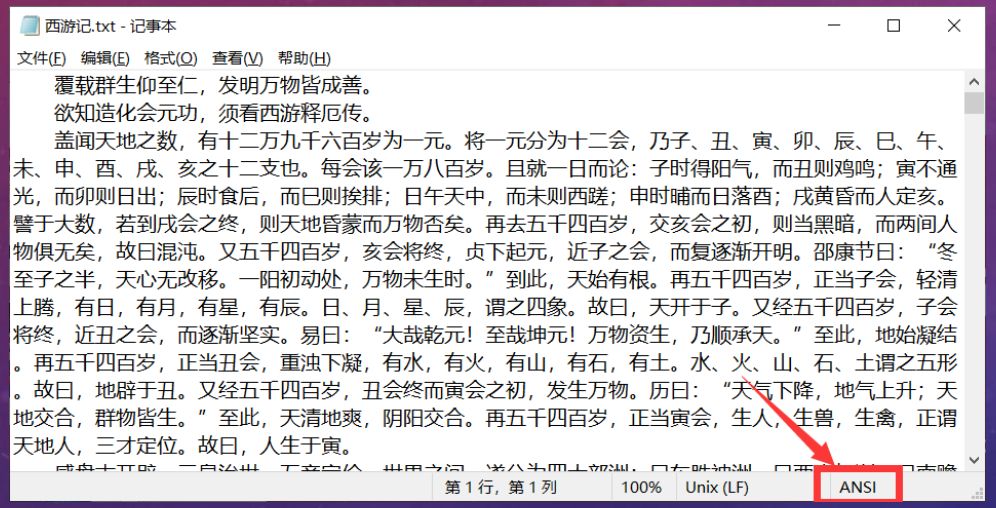
words=" ".join(xiyou)

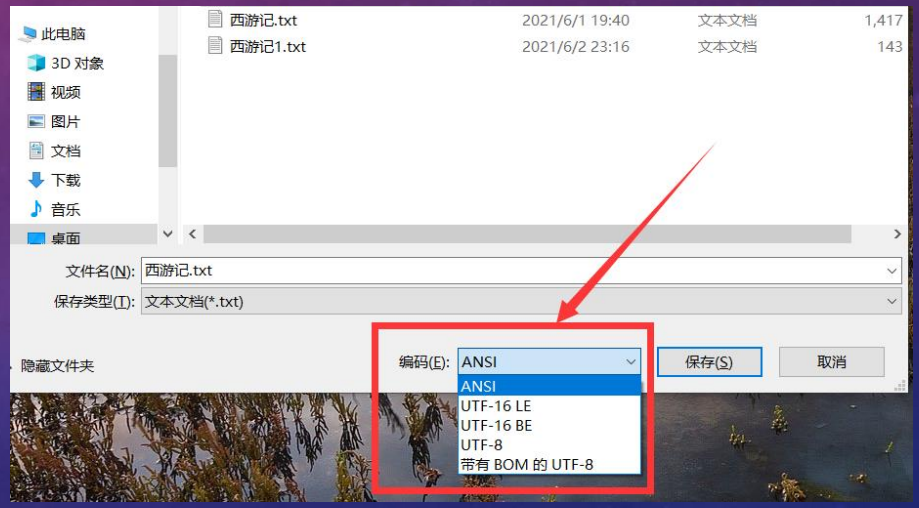
注意：学会查看文档的编码方式

小知识：学会查看修改文档的编码方式

- UTF-8、ANSI编码是字符编码的方式之一，对应着不同的编码规则，但都是将常用字符编码成对应的二进制。
- 在编写程序时，比如读取文件时我们需要注意文本的编码方式。
- 当然也可以修改文档的编码方式。

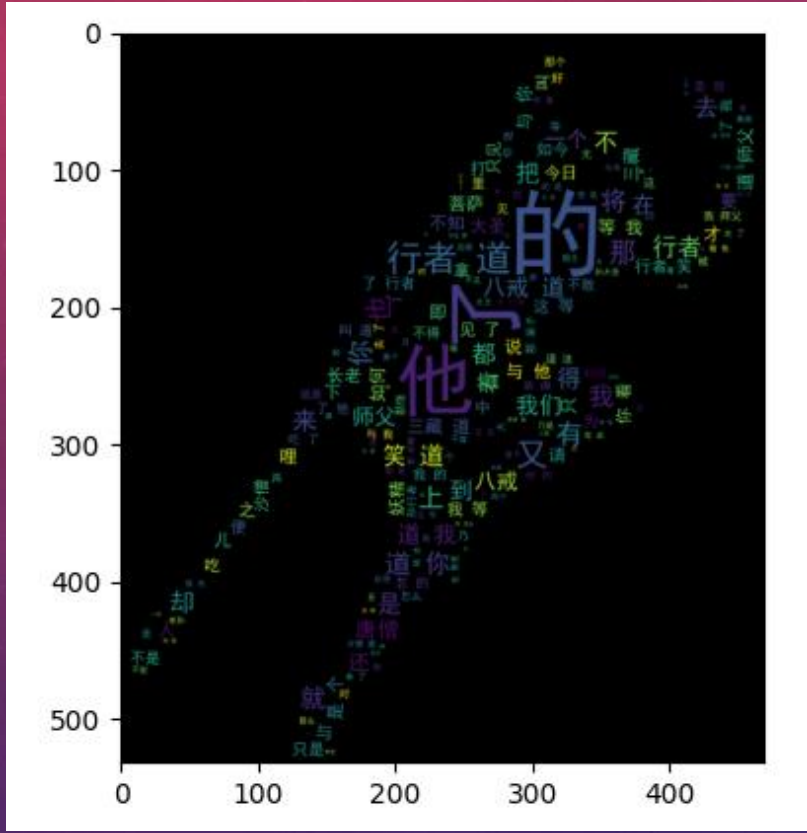
```
f1 = open("西游记.txt", "r", encoding="ANSI")
```





- 注意西游记有近60万字，数据量大，分析慢；在代码测试阶段，大家可以选择只有第一二回的西游记1.txt进行分析，分析没问题后再选择整本书籍西游记.txt进行分析。

学习活动1：对西游记文本分词并可视化

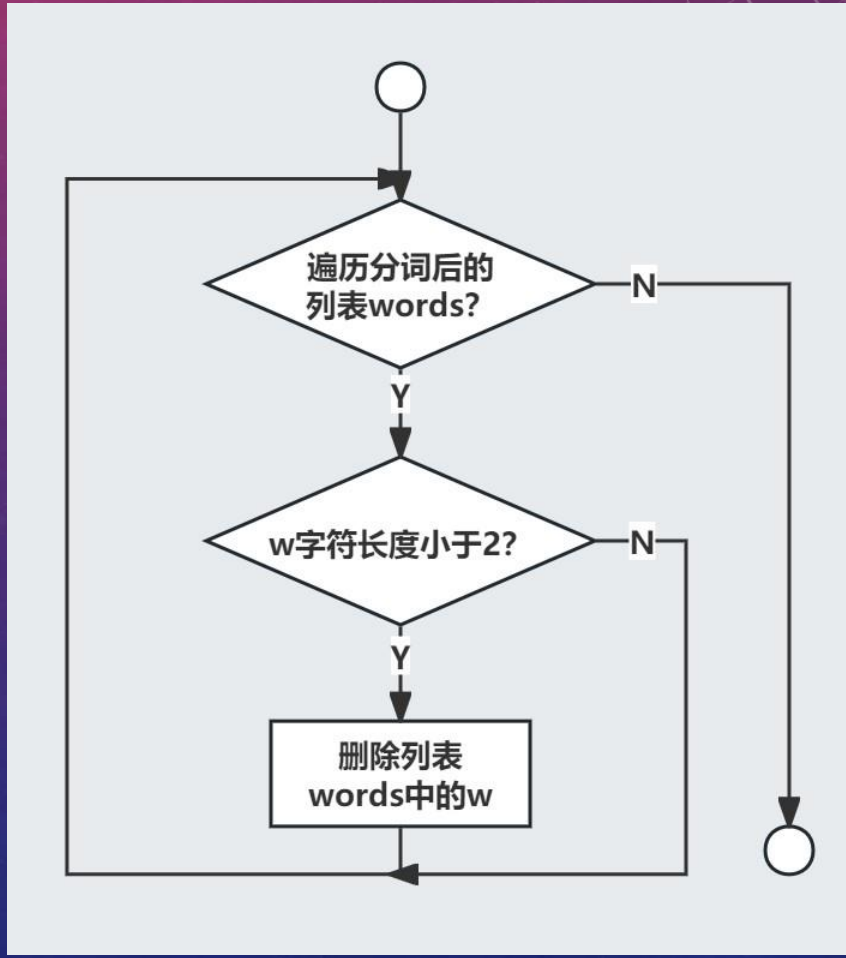



```python
#导入数据分析的外部库
import numpy as np
from PIL import Image
import matplotlib.pyplot  as plt
import wordcloud as wc
import jieba

#先将数据文件打开并读取
f1 = open("西游记.txt", "r", encoding="ANSI")
xiyou=f1.read()
f1.close()

#通过jieba分词库对中文进行分词
words = jieba.lcut(xiyou)

#通过join连接空格和列表元素
words=" ".join(words)

#形成词云
mask1 = np.array(Image.open("1.jpg"))
ciyun=wc.WordCloud(font_path="simhei.ttf",mask=mask1).generate(words)

#显示并保存词云
plt.imshow(ciyun)
plt.savefig(fname="西游记文本分析1.png")
plt.show()
```

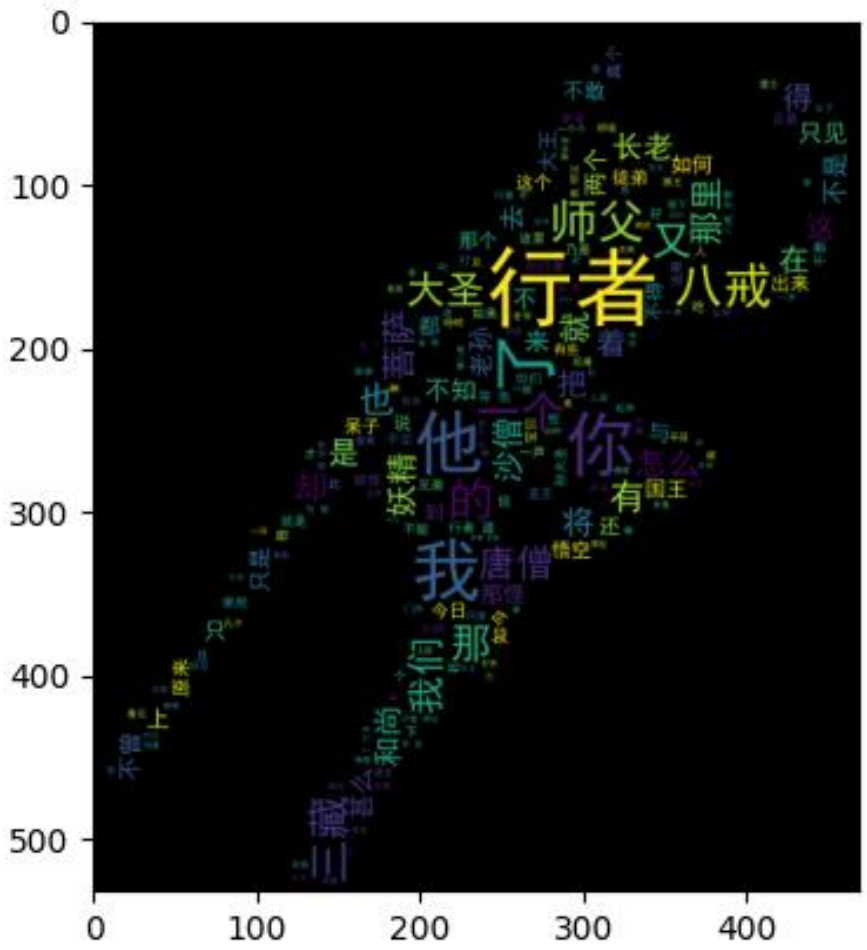
- 分词后，通过可视化分析，单个词汇最多的依次是？怎么去除单个词汇？

学习活动2：去除西游记文本中的单个词汇

- 遍历jieba分词后形成的列表words；
- 如果字符长度小于2，则删除。



```python
#导入数据分析的外部库
import numpy as np
from PIL import Image
import matplotlib.pyplot  as plt
import wordcloud as wc
import jieba

#先将数据文件打开并读取
f1 = open("西游记.txt", "r", encoding="ANSI")
xiyou=f1.read()
f1.close()

#通过jieba分词库对中文进行分词
words = jieba.lcut(xiyou)

for w in words:
    if len(w)<2:
        words.remove(w)

#通过join连接空格和列表元素
words=" ".join(words)

#形成词云
mask1 = np.array(Image.open("1.jpg"))
ciyun=wc.WordCloud(font_path="simhei.ttf",mask=mask1).generate(words)

#显示并保存词云
plt.imshow(ciyun)
plt.savefig(fname="西游记文本分析1.png")
plt.show()
```



还是会有单个词汇出现，是什么原因导致的？

循环中删除单个字符会导致列表变化，怎样修改程序让单个字符不会出现？

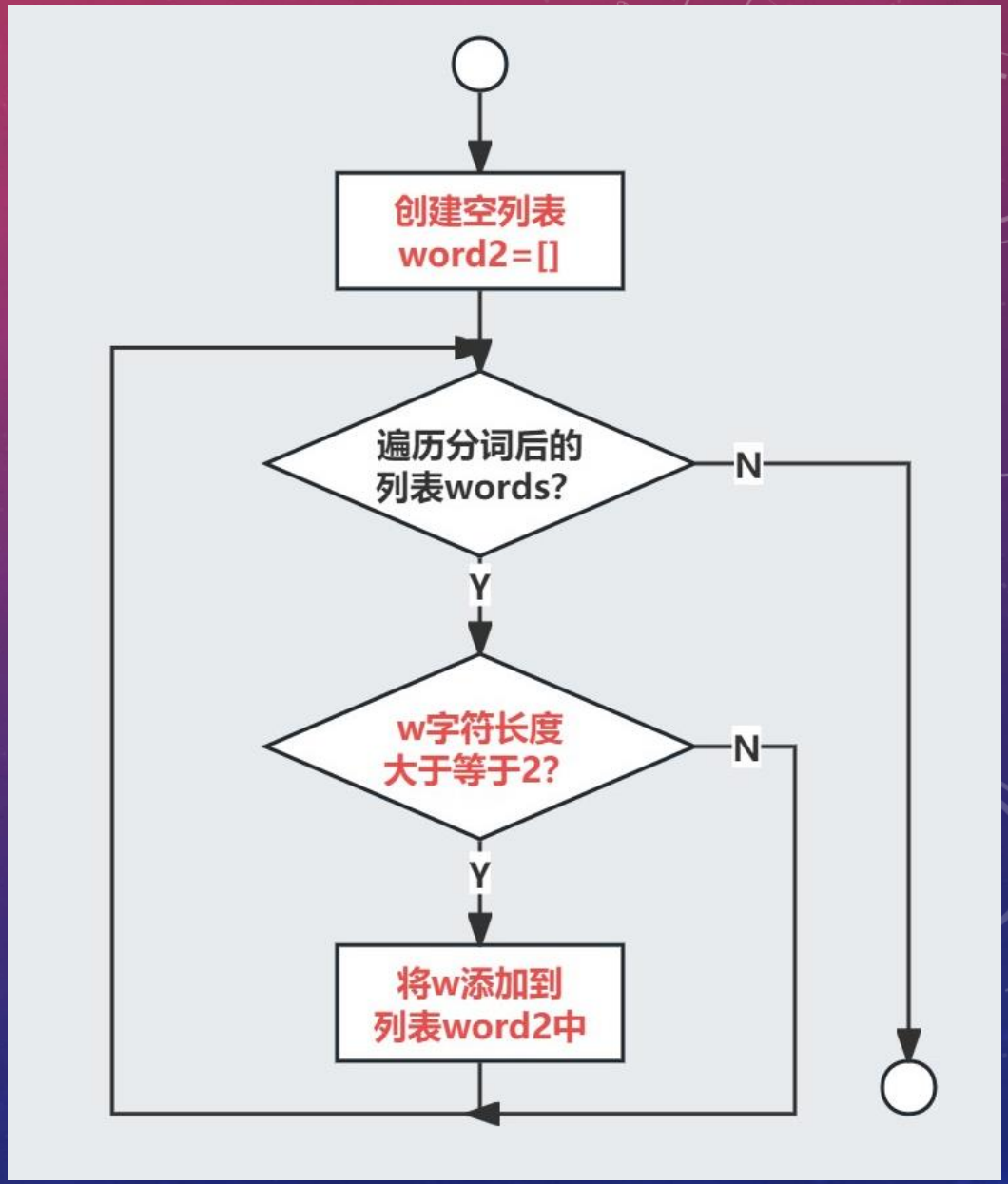
```
words=['好','歌','人','中','我']
for w in words:
    if len(w)<2:
        words.remove(w)
print(words)
```

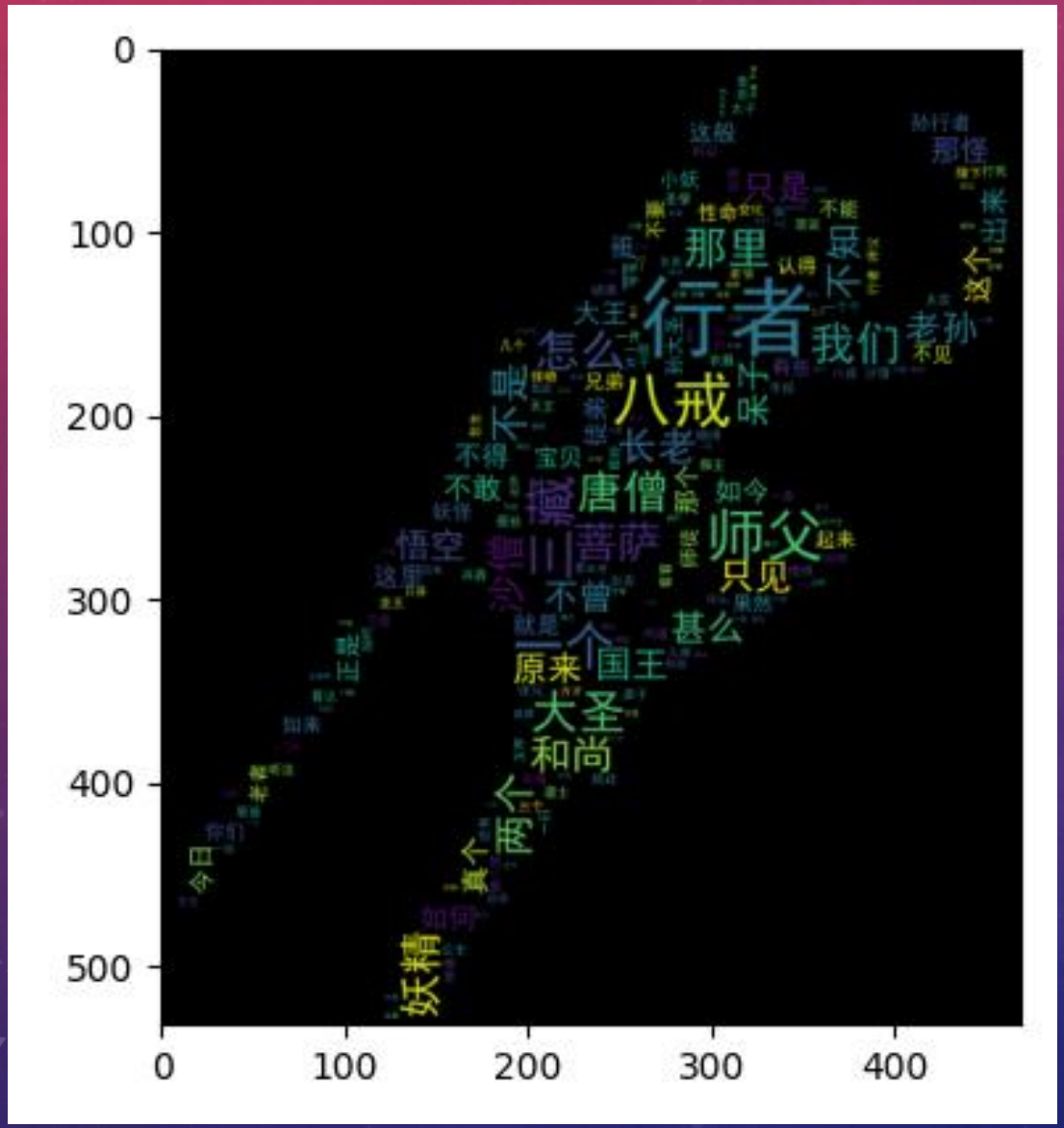
- 程序运行之后

['歌', '中']

反向思维

- 尝试将字符长度大于等于2的字符，添加到新的列表word2中，不修改原有的列表words。
  - 遍历分词后的词汇列表words；
  - 如果字符长度大于等于2，则添加到新的列表word2中。

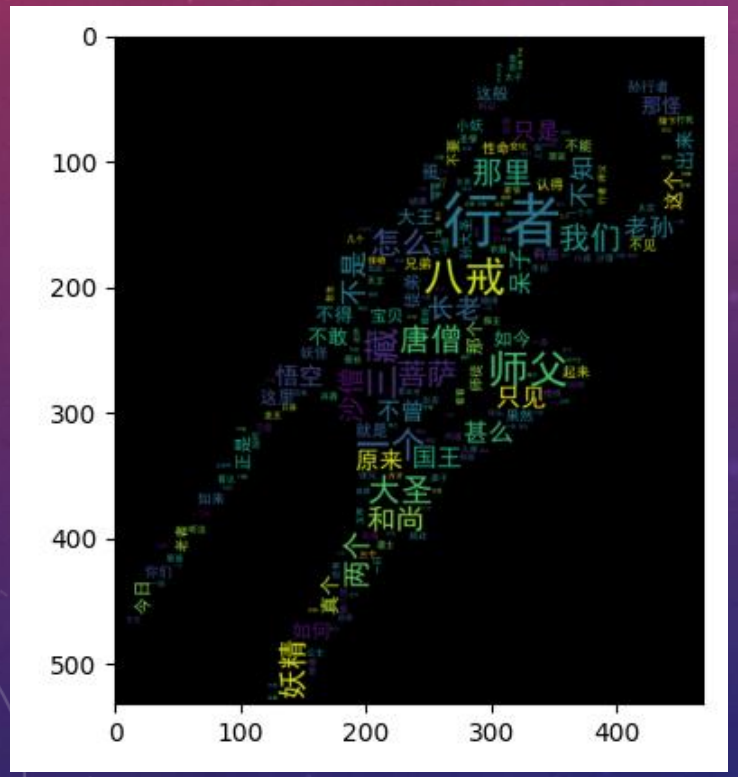




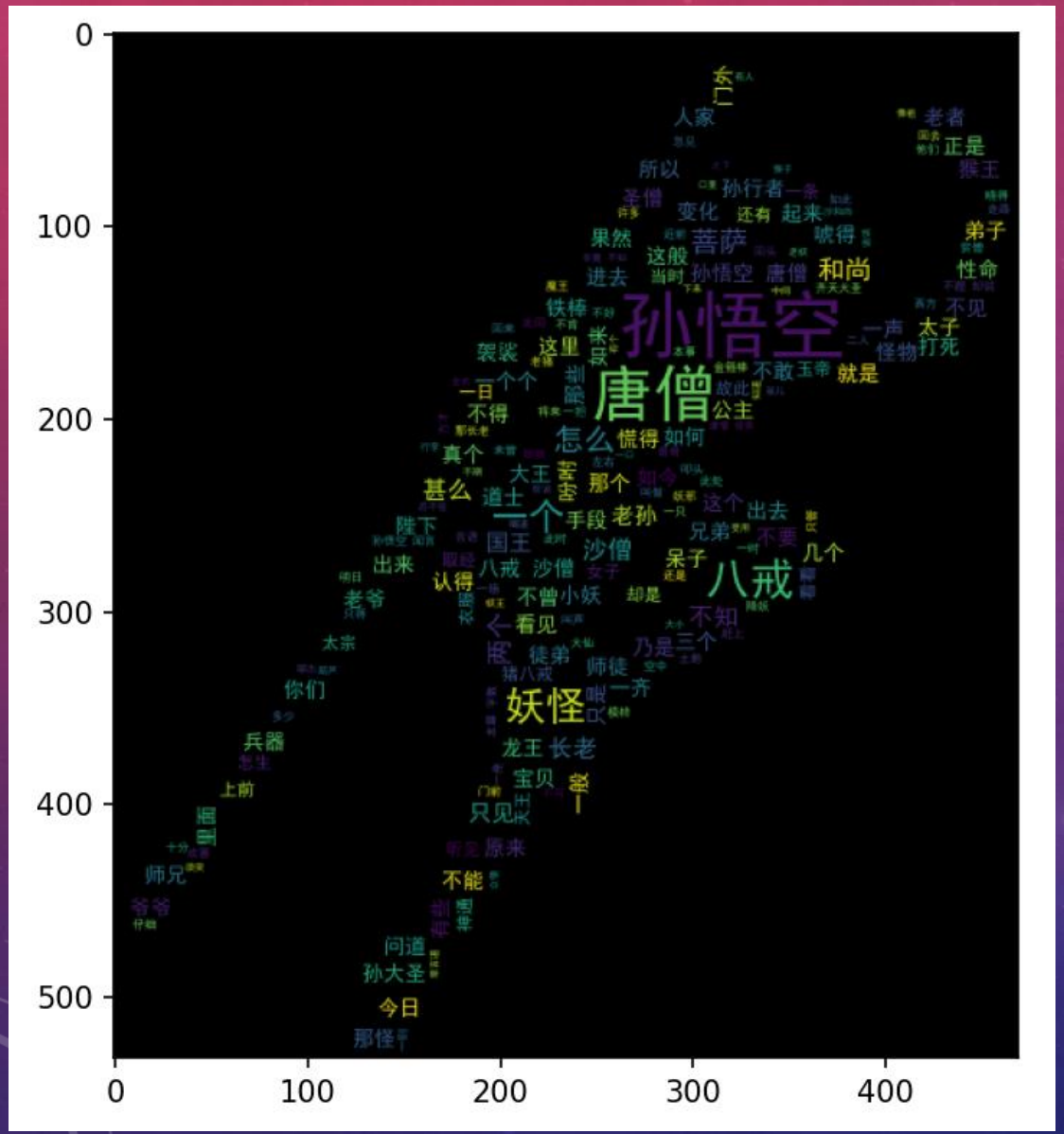

```python
#导入数据分析的外部库
import matplotlib.pyplot  as plt
import wordcloud as wc
import numpy as np
from PIL import Image
import jieba

#先将数据文件打开并读取
f1 = open("西游记.txt", "r", encoding="ANSI")
xiyou=f1.read()
f1.close()

#通过jieba分词库对中文进行分词
words = jieba.lcut(xiyou)
#遍历列表words，并将长度>=2的字符存入新列表word2
word2=[]
for w in words:
    if len(w)>=2:
        word2.append(w)
#通过join连接空格和列表元素
ds=" ".join(word2)
#形成词云
bg = np.array(Image.open("1.jpg"))
ciyun=wc.WordCloud(font_path="simhei.ttf",mask=bg)
ciyun.generate(ds)
#显示并保存词云
plt.imshow(ciyun)
plt.savefig(fname="西游记文本分析2.png")
plt.show()
```

学习活动3：部分指代及虚词及同义不同名的处理

- 有些双语词汇是中文常用的虚词、代词，如我们、不是、那里等，这些词汇不进行词云统计，应该如何处理？
- 有些词汇同义不同名，如行者、悟空、孙悟空、大圣都是指向一个人，如八戒、猪八戒都是指向一个人，如何让这些同义的词指向一个人分析。

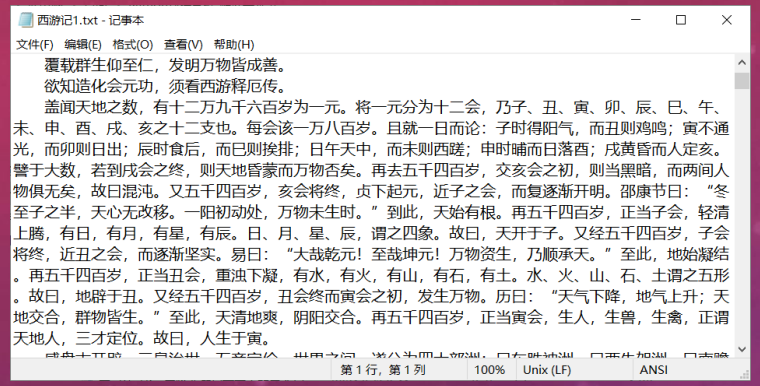


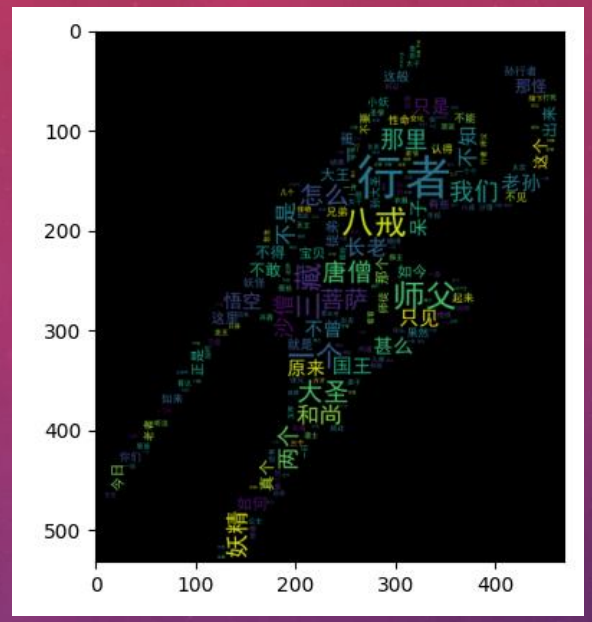

```
#通过jieba分词库对中文进行分词
words = jieba.lcut(xiyou)
#遍历列表words，并将长度>=2的字符存入新列表word2
word2=[]
for w in words:
    if len(w)>=2 and w not in ['我们','不是','那里']:
        if w=="大圣" or w=="行者" or w=="悟空" or w=="孙悟空":
            w1="孙悟空"
        elif w=="师父" or w=="三藏" or w=="唐僧":
            w1="唐僧"
        elif w=="妖精" or w=="妖魔" or w=="妖道" or w=="妖怪":
            w1="妖怪"
        else:
            w1=w
        word2.append(w1)

#通过join连接空格和列表元素
ds=" ".join(word2)
```



- 登录平台完成相应的学习评价，并将最终代码完整提交。

学习小结





数据采集

数据可视

数据整理

半结构化数据、自然语言理解、中文分词、去除冗余数据、去除无效数据、数据清洗等。

分析反馈

数据分析

数据分析

根据数据内容以及数据的可视化情况进行分析，
也可以反馈分析中的问题重新整理数据。

拓展思考：字典统计各个词汇的次数

- **遍历列表word2，当词汇出现一次则增加一次；**
- **第一次出现时，词频为1。**

- **定义空字典shu={}，键为词汇，值为对应次数；**
- **统计各个词汇的次数的关键分析：**
  - 键是词汇，所对应的值为该词汇的词频；
  - 遍历过程中，如果是第一次出现该词汇则将该词汇添加到字典中，且值为1；
  - 遍历过程中，如果词汇已经在字典中，则将原有的值增加1。

键	值
行者	100
八戒	68
唐僧	78
沙僧	43
…	…

shu={}

遍历列表：

shu[w2]=shu.get(w2,0)+1

拓展思考：字典统计各个词汇的次数

```
#先将数据文件打开并读取
f1 = open("西游记.txt", "r", encoding="ANSI")
xiyou=f1.read()
f1.close()

#通过jieba分词库对中文进行分词
words = jieba.lcut(xiyou)
#遍历列表words，并将长度>=2的字符存入列表word2
word2=[]
for w in words:
    if len(w)<2:
        continue
    else:
        word2.append(w)
#遍历列表word2，通过字典统计各个词汇的词频
shu={}
for w2 in word2:
    shu[w2]=shu.get(w2,0)+1
print(shu)

#通过join连接空格和列表元素
word2=" ".join(word2)

#形成词云
mask1 = np.array(Image.open("1.jpg"))
ciyun=wc.WordCloud(font_path="simhei.ttf",mask=mask1).generate(word2)

#显示并保存词云
plt.imshow(ciyun)
plt.savefig(fname="西游记文本分析2.png")
plt.show()
```

拓展思考：通过输入查询词汇的次数

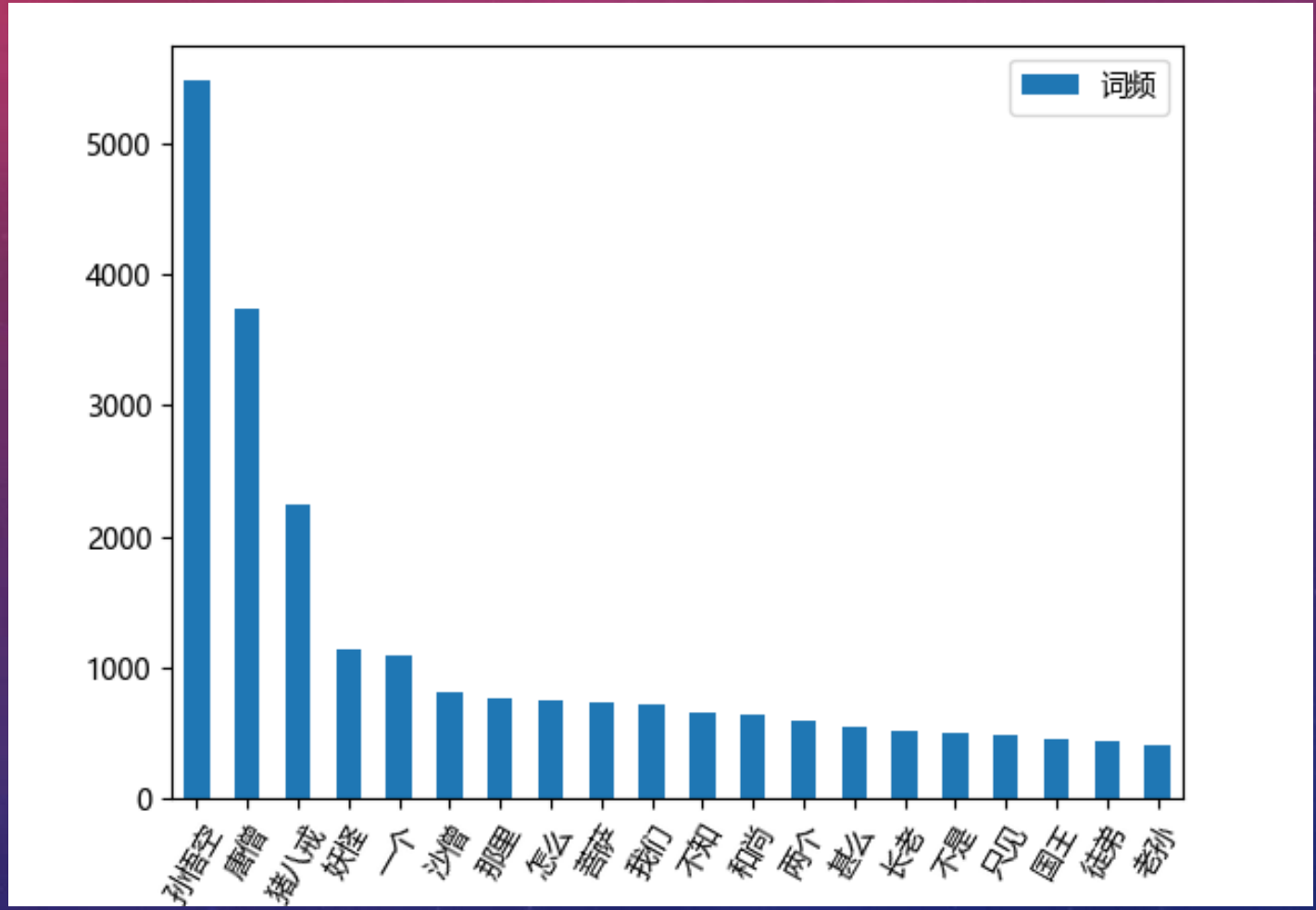
[1] 如果用户输入：退出，则退出查询循环；
[2] 如果用户输入的词汇在列表中，则输出键所对应的值；
[3] 如果用户输入的词汇不在列表中，则提示用户输入错误。

```
#通过输入查询词汇的次数
while True:
    a=input("请输入你要查询的词汇：")
    if a=="退出":
        break
    elif a in word2:
        print("出现的次数：",shu[a])
    else:
        print("词汇不在此文本中！")
```

```
请输入你要查询的词汇：大圣
出现的次数： 889
请输入你要查询的词汇：行者
出现的次数： 4078
请输入你要查询的词汇：唐僧
出现的次数： 802
请输入你要查询的词汇：八戒
出现的次数： 1677
请输入你要查询的词汇：沙僧
出现的次数： 721
请输入你要查询的词汇：妖怪
出现的次数： 215
请输入你要查询的词汇：happy
词汇不在此文本中！
请输入你要查询的词汇：退出
```

拓展思考：
如何将数据按照一定顺序排列并通过柱形图显示出来？

词语	词频
孙悟空	5472
唐僧	3730
猪八戒	2239
妖怪	1139
一个	1089
沙僧	815
那里	767
怎么	754
菩萨	730
我们	725
不知	657
和尚	644
两个	594
甚么	551
长老	512
不是	507
只见	485
国王	456
徒弟	439
老孙	408



基于Python可视化分析《西游记》文本

谢谢聆听

昆明市第一中学西山学校 齐洪