

从词云可视化看数据处理过程

——以云南词条可视化分析为例

昆明市第一中学西山学校 齐洪

数据处理过程



数据收集的方式:

实验数据采集
调查研究采集
网络数据采集
系统日志采集
传感器设备采集
.....

数据整理的内容:

数据缺失
数据重复
数据格式不一致
数据逻辑错误
.....

数据分析的工具:

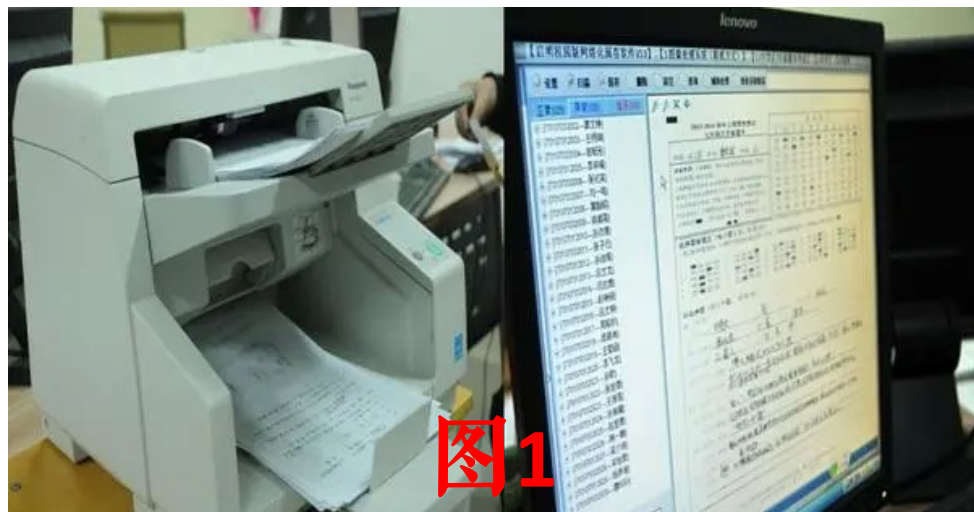
Excel数据分析
• If、Sumif()、
averageif()、rank()等
Python数据分析
• numpy、pandas等库
.....

数据可视化:

Excel数据可视化
Python数据可视化
• matplotlib、wordcloud等
外部库

线性图、散点图、柱状图、
词云图等
.....

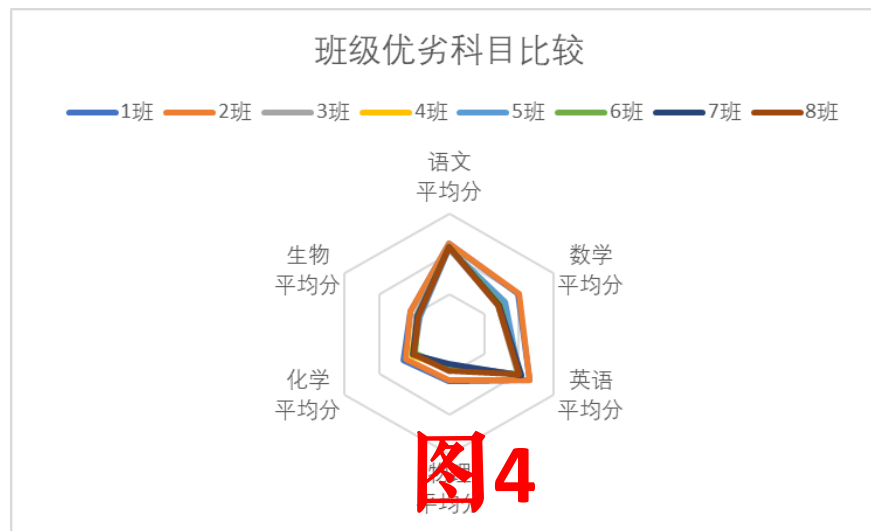
数据分析处理过程——以Excel为例



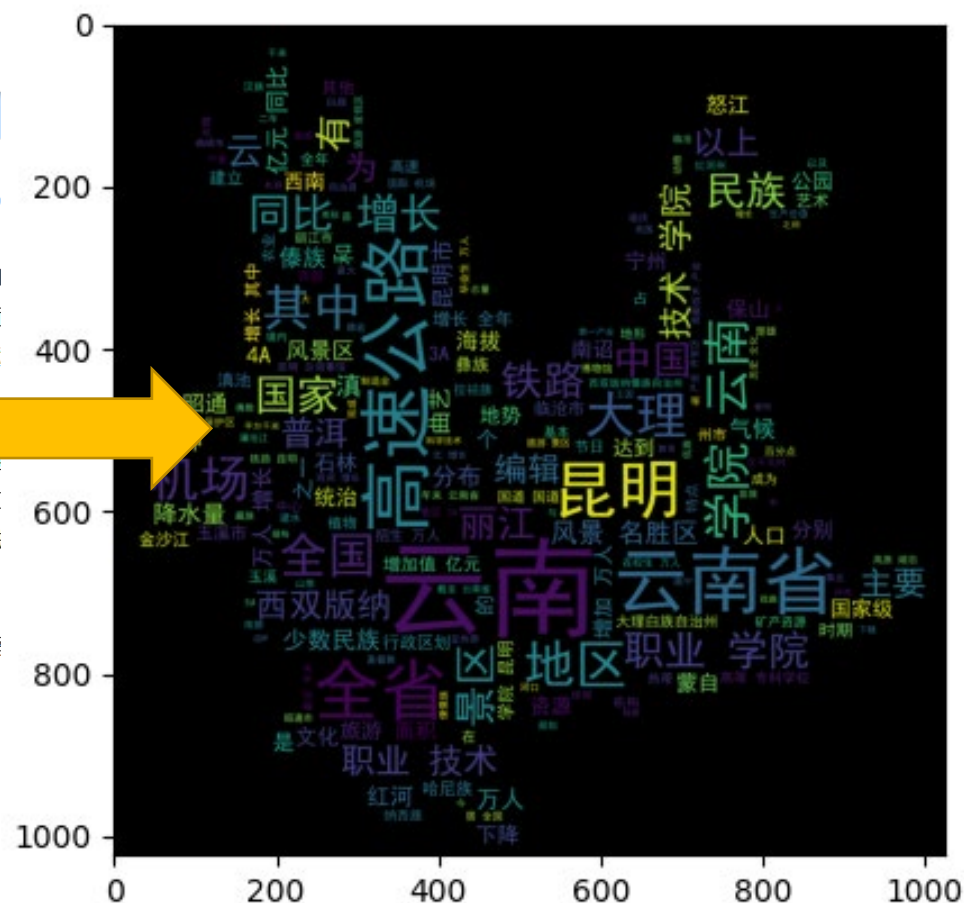
准考证号	班级	姓名	语文					得分	校次
			得分	校次	校次进退步	班次	班次进退步		
17324743	高三年级1班	李雨橙	120	15	264	2	32	127	3
17312817	高三年级1班	杜磊	118	39	274	4	33	106	49
17315915	高三年级2班	林金鑫	117	57	-23	10	-1	115	18
17324838	高三年级1班	董文韵	118	39	152	4	19	134	1
17316624	高三年级1班	袁林	116	79	234	13	24	113	26
17325141	高三年级2班	陈芝杰	116	79	377	13	31	117	12
17315908	高三年级2班	刘蕾	122	5	0	2	-1	118	9
17315933	高三年级2班	刘鑫宇	112	187	288	27	19	126	4
17315913	高三年级2班	何恩进	115	101	-89	15	-12	119	7
17325140	高三年级3班	陈明	115	101	17	8	4	70	317
17321641	高三年级7班	李佳懿	121	11	10	2	1	58	439
17313946	高三年级5班	程歆倬	116	79	200	4	26	62	394
17321629	高三年级7班	唐瑞泉	未扫	156	79	16	9	40	529
17324814	高三年级10班	李卓亭	89	553	-160	35	-10	51	484
17325133	高三年级3班	顾皓元	112	187	48	15	10	70	317
17317450	高三年级5班	史卓灵	105	391	-112	29	1	63	388
17314658	高三年级6班	梁伦博	112	153	15	15	-13	未扫	338
17315184	高三年级4班	王宇豪	115	101	22	3	24	55	459
17324828	高三年级8班	李光敏	未扫	151	48	46	-16	80	231
17325143	高三年级3班	付堯	106	370	139	24	27	75	268

某学校高中2022年理科期中测评											期中考试成绩统计分析										
1	班级	姓名	语文	数学	英语	物理	化学	生物	总分	排名	1	班级	语文平均分	数学平均分	英语平均分	物理平均分	化学平均分	生物平均分	总分平均分		
2	1班	王明玉	120	127	120	71	68	68	574	1	2	1班	112.50	98.62	115.92	58.40	65.38	56.30	507.12		
3	1班	李雷	118	106	122.5	67	81	77	571.5	2	3	2班	112.62	100.04	115.90	57.36	62.80	56.42	505.14		
4	2班	张琳	117	115	122	69	80	66	569	3	4	3班	106.51	70.25	99.30	38.13	52.12	47.13	409.49		
5	1班	刘明	118	134	124.75	61	71	55	563.8	4	5	4班	107.44	75.73	100.85	42.15	53.48	45.54	425.18		
6	1班	王鑫	116	113	126.5	78	66	64	563.5	5	6	5班	106.36	79.84	100.49	40.49	50.49	43.08	420.75		
7	2班	陈宏	116	117	122.5	64	79	63	561.5	6	7	6班	108.79	73.10	100.48	42.17	49.17	44.77	416.21		
8	2班	张东	122	118	129	52	75	64	560	7	8	7班	108.21	70.98	103.24	37.27	51.88	45.25	416.17		
9	2班	李进	112	126	128.5	71	65	56	558.5	8		8班	107.47	70.04	99.43	46.37	50.83	43.72	416.87		
10	2班	王鑫	115	119	117	65	75	67	558	9											
11	1班	杨冲	118	114	119.5	75	68	61	555.5	10											
12	2班	杨红	118	116	124.25	76	71	50	555.3	11											
13	2班	赵思	115	110	132	62	81	55	555	12											
14	2班	王昊喆	115	112	112	78	74	63	554	13											
15	1班	杨通	115	107	125.5	59	76	66	548.5	14											
16	2班	杨智	113	119	114.5	70	70	60	546	15											
17	2班	李懿书	117	120	124.5	66	65	53	545	16											
18	2班	满鑫	113	116	103.5	73	70	67	542.5	17											
19	1班	王思涵	116	115	114	63	79	53	540	18											
20	2班	严宇若	112	98	134	70	67	59	540	18											

图3

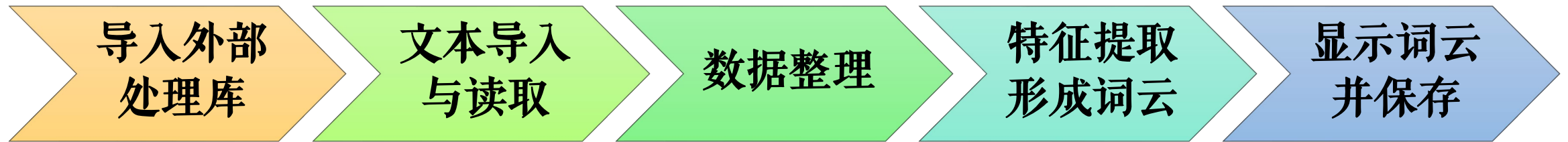


对“云南”百度百科词条的可视化分析



数据采集：网络采集

Python处理文本数据及词云可视化过程



```
#导入外部库
#matplotlib.pyplot是绘图库
#wordcloud是词云制作库
import matplotlib.pyplot as plt
import wordcloud as wc

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

```
#形成词云
a=wc.WordCloud(font_path="simhei.ttf")
a.generate(fs)

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

提取词频制作词云（词频越大词云越突出）

#导入外部库, matplotlib.pyplot绘图库

#wordcloud是词云制作库

```
import matplotlib.pyplot as plt
```

```
import wordcloud as wc
```

#先将数据文件打开并读取

```
f=open("yunnan.txt",encoding="utf-8")
```

```
fs=f.read()
```

```
f.close()
```

#形成词云

```
a=wc.WordCloud(font_path="simhei.ttf")
```

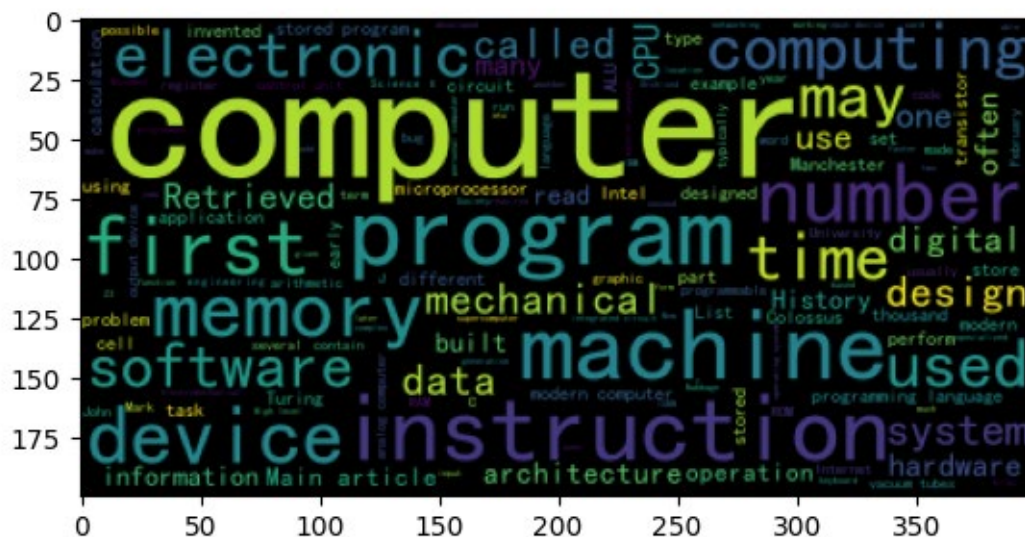
```
a.generate(fs)
```

#显示并保存词云

```
plt.imshow(a)
```

```
plt.savefig(fname="云南词条分析1.png")
```

```
plt.show()
```



- **英文词云和中文词云能否直接反应文本主旨？**
- **英文和中文形成词云中，怎么提取词汇并统计词频？（空格及标点符号）**
- **下面这句中文的词频：**
- **14个云南省级历史文化名村和1个云南省级历史文化街区**

- 对于中文词云图的设计，一般先进行分词、连接等数据处理，再经过wordcloud库进行特征提取并形成词云图。

yunnan1.txt 14个云南省级历史文化名村和1个云南省级历史文化街区

#打开文件并读取文本数据

```
f=open("yunnan1.txt",encoding="utf-8")
```

```
fs=f.read()
```

```
f.close()
```

```
print(fs)                                  14个云南省级历史文化名村和1个云南省级历史文化街区
```

#通过jieba分词库对中文进行分词

```
ws=jieba.lcut(fs)                        ['14', '个', '云南', '省级', '历史', '文化', '名村', '和', '1', '个',
```

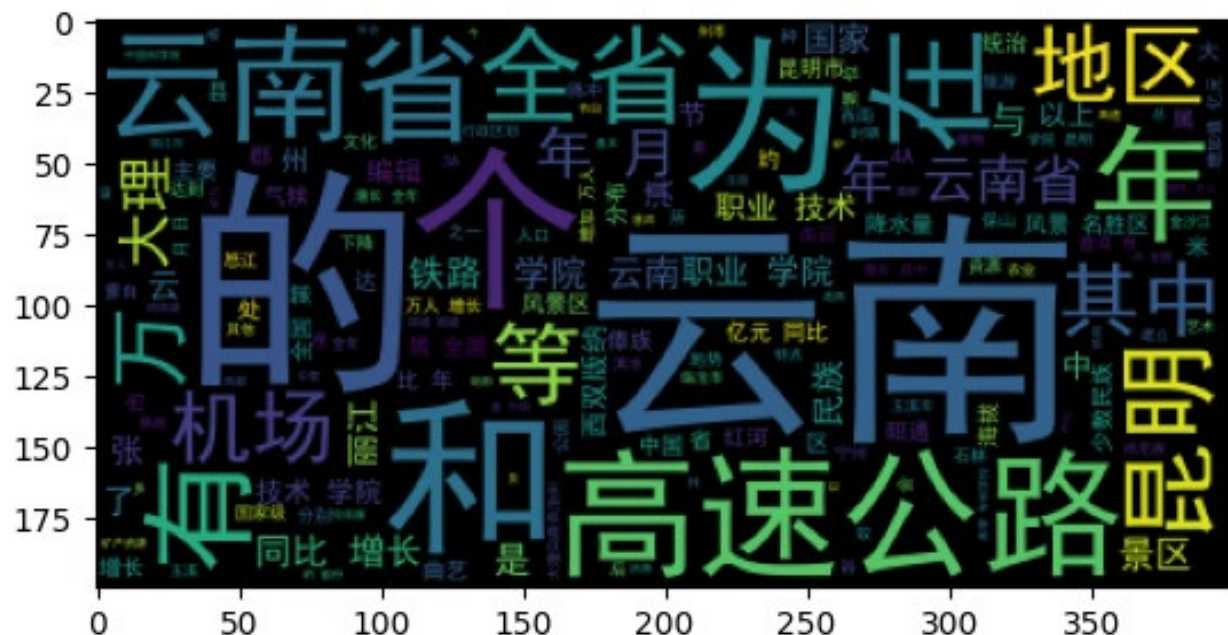
```
print(ws)                                '云南', '省级', '历史', '文化', '街区']
```

#通过join方法 用空格将列表各元素连接

```
ds=" ".join(ws)                        14 个 云南 省级 历史 文化 名村 和 1 个 云南 省级 历
```

```
print(ds)                               史 文化 街区
```


中文分词——jieba库



- jieba.lcut()产生的数据类型是什么类型?
- 现在有没有很好地反映这篇文章的特点? 什么原因导致的?

#导入外部库，matplotlib.pyplot是绘图库

#wordcloud是词云制作库

```
import matplotlib.pyplot as plt
```

```
import wordcloud as wc
```

```
import jieba
```

#打开文件并读取文本数据

```
f=open("yunnan.txt",encoding="utf-8")
```

```
fs=f.read()
```

```
f.close()
```

#对数据进行整理

#通过jieba分词库对中文进行分词

```
ws=jieba.lcut(fs)
```

#通过join方法 用空格将列表各元素连接

```
ds=" ".join(ws)
```

#提取特征 形成词云

```
a=wc.WordCloud(font_path="simhei.ttf")
```

```
a.generate(ds)
```

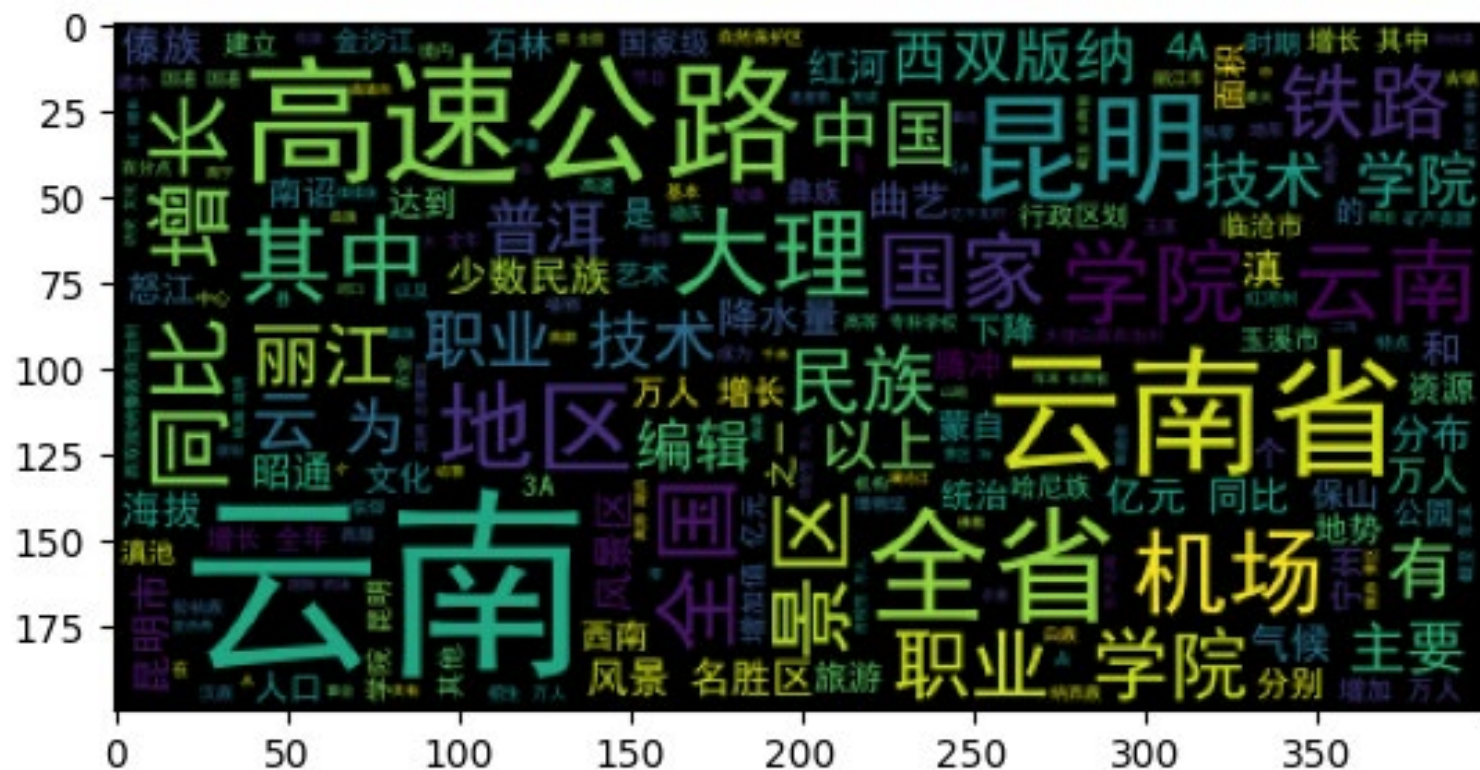
#显示并保存词云

```
plt.imshow(a)
```

```
plt.savefig(fname="云南词条分析1.png")
```

```
plt.show()
```

数据处理：去除单个字符



- 还有单个字符没有？什么原因导致的？
- 假设ws为["云南','的','个','昆明']

```
#导入外部库，matplotlib.pyplot是绘图库
#wordcloud是词云制作库
import matplotlib.pyplot as plt
import wordcloud as wc
import jieba
#打开文件并读取文本数据
f=open("yunnan.txt",encoding="utf-8")
fs=f.read()
f.close()
```

```
#数据整理，通过jieba分词库对中文进行分词
ws=jieba.lcut(fs)
```

```
#循环中将单个字符删除
```

```
for w in ws:
    if len(w) == 1:
        ws.remove(w)
```

```
#通过join方法 用空格将列表各元素连接
ds=" ".join(ws)
```

```
#形成词云
```

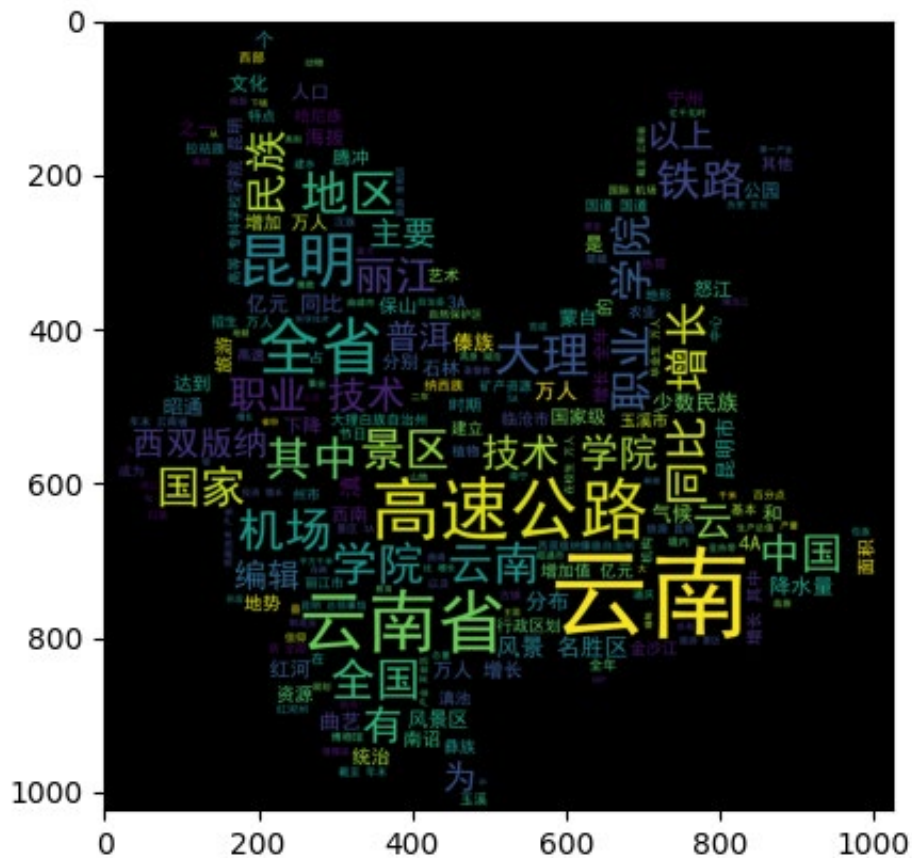
```
a=wc.WordCloud(font_path="simhei.ttf")
a.generate(ds)
```

```
#显示并保存词云
```

```
plt.imshow(a)
plt.savefig(fname="云南词条分析1.png")
plt.show()
```

可视化修改： 用云南图片作为形状

- 词云的形状不是矩形，设计成其它形状，比如用云南的地图形状作为词云的形状？代码中对图片进行怎样处理？



#导入外部库，matplotlib.pyplot是绘图库

#wordcloud是词云制作库

```
import matplotlib.pyplot as plt
```

```
import wordcloud as wc
```

```
import jieba
```

```
import numpy as np
```

```
from PIL import Image
```

#打开文件并读取文本数据

```
f=open("yunnan.txt",encoding="utf-8")
```

```
fs=f.read()
```

```
f.close()
```

#数据整理，通过jieba分词库对中文进行分词

```
ws=jieba.lcut(fs)
```

#循环中将单个字符删除

```
for w in ws:
```

```
if len(w) == 1:
```

```
ws.remove(w)
```

#通过join方法 用空格将列表各元素连接

```
ds=" ".join(ws)
```

#形成词云

```
bg = np.array(Image.open("yunnan.jpg"))
```

```
a=wc.WordCloud(font_path="simhei.ttf",mask=bg)
```

```
a.generate(ds)
```

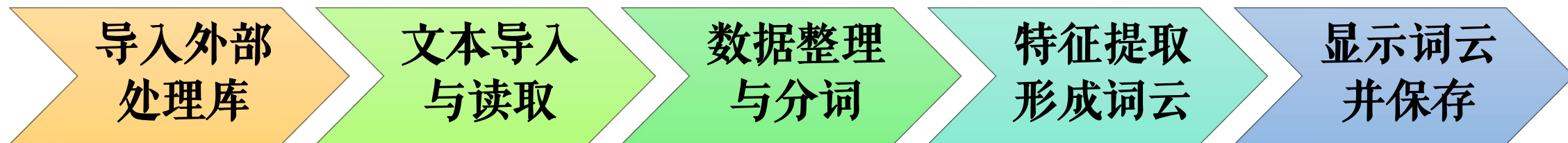
#显示并保存词云

```
plt.imshow(a)
```

```
plt.savefig(fname="云南词条分析1.png")
```

```
plt.show()
```


Python处理数据及词云可视化的过程



- 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