**数据预处理**

**准备jieba中文分词库**

* 通过pip install jieba安装
* 通过import jieba引用

**对源文件夹SogouQ中的每个文件提取搜索内容并进行分词，然后保存进一个输出文件data中**

* 获取文件夹SogouQ下所有文件名，据此遍历处理所有文件

|  |
| --- |
| Python file\_names = os.listdir(file\_base) |

* 文件的编码未知，用utf-8会乱码，而根据文件中有很多中文，因此通过gbk解码，实测效果不错

|  |
| --- |
| Python with open(file\_path, mode='r', encoding=file\_encoding, errors="ignore") as fin:  lines = list(line for line in fin.read().split('\n') if line != "") |

* 提取每行搜索内容部分

|  |
| --- |
| Python title = re.sub(r'(?:\s+)', ' ', line.strip()).split(' ')[2][1:-1] |

效果举例：

|  |
| --- |
| Python 00:00:00 8561366108033201 [汶川地震原因] 3 2 www.big38.net/ => 汶川地震原因 |

* 使用jieba进行分词，并剔除虚词、连词等词性的影响

|  |
| --- |
| Python words = pseg.cut(str(title)) split = [] for word, flag in words:  if flag in ['m', 'q', 'r', 'p', 'c', 'u', 'xc', 'w', 'x']: continue  split.append(word) |

效果举例：

|  |
| --- |
| Python 汶川地震原因 => 汶川 地震 原因 |

* 去除每个搜索里重复的搜索词并将结果追加入data文件

|  |
| --- |
| Python if len(split) > 0:  split = list(set(split))  fout.write(' '.join(split) + '\n') |

**源码**

|  |
| --- |
| Python import re import os  import jieba.posseg as pseg  file\_base = 'D:/college/junior/semester1/bigdata/pj/SogouQ/' out\_base = 'D:/college/junior/semester1/bigdata/pj/output/' # file\_base = 'in/' # out\_base = 'out/'  file\_encoding = 'gbk' out\_encoding = 'utf-8'  def preprocess():  file\_names = os.listdir(file\_base)  for file\_name in file\_names:  file\_path = file\_base + file\_name  out\_path = out\_base + 'data'  with open(file\_path, mode='r', encoding=file\_encoding, errors="ignore") as fin:  lines = list(line for line in fin.read().split('\n') if line != "")  with open(out\_path, mode='a', encoding=out\_encoding) as fout:  for line in lines:  title = re.sub(r'(?:\s+)', ' ', line.strip()).split(' ')[2][1:-1]  words = pseg.cut(str(title))  split = []  for word, flag in words:  if flag in ['m', 'q', 'r', 'p', 'c', 'u', 'xc', 'w', 'x']: continue  split.append(word)  if len(split) > 0:  split = list(set(split))  fout.write(' '.join(split) + '\n')  print(f'complete {file\_path}')  if \_\_name\_\_ == '\_\_main\_\_':  preprocess() |

**运行**

* 原数据共有28个文件，总大小为4.23G，本机运行上述代码3h后，数据预处理完毕，合并输出为1个文件，大小约为955MB，大约共有51 000 000条数据