QINMIN HU

2018年4月10日

LAB 2 构建倒排索引

一. 实验目的

构建倒排索引:

- 对每个块独立建立倒排索引
- 将所有的独立索引进行合并

二. 实验步骤

1. 对每个块独立建立倒排索引

```
def create_inverted_index():
    test={}
    i=0
    for words in text:
        for word in words.split():
            if word==';' or word==',' or word=='$' or word=='//' or word=='/' or word=='{' or word=='}':
                （建立字典之前，去除一些没有意义不需要添加到字典中的符号）
                continue
            if word in dict_test:
                if docno[i] in dict_test[word]:
                    dict_test[word][docno[i]]=dict_test[word][docno[i]]+1
                else:
                    dict_test[word][docno[i]]=1
            else:
                test[docno[i]]=1
                dict_test[word]=test
                test={}
        i=i+1
```

（将出现的词语构建字典，字典中key为出现词语，value为一个内部字典，内部字典为此词语出现的文本，key为出现文本名，value为此词语出现在此文本中的次数）

2. 将所有的独立索引进行合并&写入inverted_index.json文件

```
def add_invert():
    dict_final=my_dict[0]
    for word in my_dict[1:]:
        for key1 in word:
            if key1 in dict_final:
                dict_final[key1]=dict(Counter(dict_final[key1])+Counter(word[key1]))
            else:
                dict_final[key1]=word[key1]
```

（对已经建立好的倒排索引进行合并，至dict_offi字典中）

```
    dict_sorted=sorted(dict_final.items(),key=lambda d:d[0])
```

（对已经合并的倒排索引dict_offi排序）

```
    file_write=open("inverted_index.json",'w')
    i=0
    for word in dict_sorted:
        jsobj=json.dumps(word[1])
        if i==0:
            file_write.write('{ "'+word[0]+'" :'+jsobj+',')
        else:
            file_write.write('"'+word[0]+'" :'+jsobj+',')
        i=i+1
    file_write.write('}')
    file_write.close()
```

（将排好序的dict_write写入json文件）