《人工智能软件开发与实践》

(2023 学年 秋季 学期)

作

业

报

告

 学
 号:

 姓
 名:

 班
 级:

 任课教师:

实验名称: 名实体频度统计

成绩:

一、实验目的

熟悉字符串编程的有关内容,主要包括文件读入、字符串切分、词频统计技术。统计如下 实体类型出现的数量以及所有实体出现次数最多的前 10 个。

符号	名称
nr	人名
ns	地名
nt	机构名
nx	外文字符
nz	其他专名

二、实验内容

- 1. 打开文件 199801. txt。
- 2. 删除每行开始的句子标识。
- 3. 统计每行中出现的所有名实体,需要注意以下特殊情况:
 - 语料中人名的姓和名是分开标准的,统计时应该作为一个实体;
 - 江/nr 泽民/nr 算一个实体
 - 李/nr 鹏/nr 算一个实体
 - 嵌套实体用[]括起来,即使里面还有粒度更小的实体,也按长度最大的实体计算,只算一次,不重复计算。

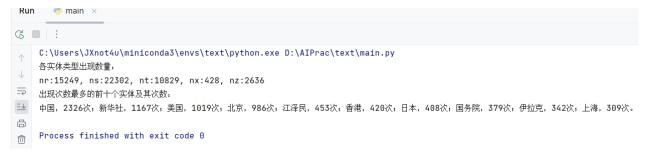
如:[中国/ns 香港/ns 特别/a 行政区/n]ns 整体算一个,里面的实体不算

- 4. 在整个文件中统计出各种名实体类型出现的次数,以及所有实体出现次数最多的前 10 个(同时输出出现次数)。
- 三、使用的算法名称(若无,可以不填)
- 四、程序源码(拷贝至此处,同时作为附件和报告再一份单独的程序)

```
f = open("199801.txt", "r", encoding='ANSI') # 打开文件
# 变量
dic = {} # 用字典 dic 统计各实体出现次数
           # 各类型计数
nr count = 0
ns count = 0
nt count = 0
nx count = 0
nz count = 0
nr flag = 0 # 人名标志,用于判断合并中文姓名
pa flag = 0 # 括号标志,用于判断实体是否在括号中
entity = "" # 存储实体名
for line in f:
  if len(line) != 0: # 行非空
     words = line.split()[1:] # 以空格分词为列表 words 并去掉句子标识
  else: # 跳过空行
     continue
  for word in words:
     tag = word.split('/')[1] # 获得词的类型
                      # 连续两个nr, 在处理前一个词时被合并为一个中文名, 跳过此轮循环
     if nr flag == 1:
        nr flag = 0
        continue
     if word[0] == "[": # 括号开始
        pa flag = 1
        entity = word.split('/')[0][1:] # 获取嵌套实体的第一个实体名并去掉括号
11 //
        continue
     if pa flag == 1: # 词在括号中,将其拼接进实体名
        entity = entity + word.split('/')[0]
        if "]" in word: # 括号结束
           tag = word.split(']')[1]
        else:
           continue
     if tag == "nr": # 处理人名, 若后一个也是人名则拼接
        if words.index(word) != len(words) - 1:
           nextWord = words[words.index(word) + 1]
           if nextWord.split('/')[1] == "nr":
              nr flag = 1
              nr count += 1
              entity = word.split('/')[0] + nextWord.split('/')[0]
        else:
           nr count += 1
```

```
# 计数
      if tag == "ns":
         ns count += 1
      if tag == "nt":
         nt count += 1
      if tag == "nx":
         nx count += 1
      if tag == "nz":
         nz count += 1
      # 非实体
      if tag != "nr" and tag != "ns" and tag != "nt" and tag != "nx" and tag !=
"nz":
         continue
      # 处理人名和括号的特殊情况: 实体名已拼接完毕
      if pa flag == 1:
         pa_flag = 0
      else: # 非特殊情况,直接获取实体名
         if not nr flag == 1:
            entity = word.split('/')[0]
      # 排除实体名为空
      if len(entity) == 0:
         continue
      else:
         if entity in dic:
            dic[entity] += 1
         else:
            dic[entity] = 1
         entity = ""
                     # 清空实体名
# 将字典转为二元组列表进行排序
dic sorted = sorted(list(zip(dic.values(), dic.keys())), reverse=True)
# 结果输出
print ("各实体类型出现数量:")
print("nr:%d, ns:%d, nt:%d, nx:%d, nz:%d" % (nr_count, ns_count, nt_count, nx_count,
nz count))
print ("出现次数最多的前十个实体及其次数:")
for i in range(9):
   print(list(dic sorted)[i][1] + ", %d次; " % list(dic sorted)[i][0], end='')
print(list(dic_sorted)[9][1] + ", %d次。" % list(dic sorted)[9][0])
```

五、程序运行结果(将程序运行结果的截图,拷贝至此处)



六、心得体会和遇到的困难

对 python 语言依旧比较陌生,特殊情况处理起来有些棘手,要简化逻辑,善于上网搜索解决问题,并熟练运用 debug 发现并解决问题。