Homework 1

2019 年 10 月 30 日

In [1]:
```
import re

f = open('hw1_news.txt')
# 用一个 flag 变量来表示目前是哪一篇新闻，初始值为 1
flag = 1
# 利用嵌套字典来进行存储，{news_flag:{word:frequency}}
name_freq = {}
while True:
    line = f.readline()
    if not line:
        break
#     判断到底是第几篇新闻
#     如果遇到标题则更新 flag 的值
#     否则 flag 保持原来的值
    if re.search('新闻 2',line):
        flag = 2
    if re.search('新闻 3',line):
        flag = 3
#     方法 2
#     if re.search('新闻',line):
#         flag = re.findall('新闻 (\d)',line)[0]

#     文中人名的规律：若干个英文字母 + 空格 + 可能存在的 middle name（包括.）+ 空格 + 若干个英文字母（姓）
    if re.search('[a-zA-Z \.]+',line):
#         如果这一行字符串中存在人名
#         则用 findall 方法来匹配出这一行所有的人名
#         方法 1
        names = re.findall('([a-zA-Z \.]+)',line)
#         方法 2
#         names = re.findall('(([a-zA-Z\.]+ ){1,2}[a-zA-Z\.]+)',line)
#         使用方法 2 的话，后面的语句需要对应调整
#         返回的结果是一个 n 行 2 列的嵌套列表
#         第一个元素为外面的括号提取的文本（人名）
#         第二个元素为里面的括号提取的文本（given names）
        if flag not in name_freq:
#             如果还没有属于这一篇新闻的字典，则添加一个
            name_freq[flag] = {}
        for n in names:
#             对匹配成功的人名列表进行遍历
            if n not in name_freq[flag]:
#                 如果字典中没有这个人名，则添加这个人名
                name_freq[flag][n] = 1
#                 如果是方法 2 进行匹配，则应该为 name_freq[flag][n[0]] = 1
```

```
                else:
        #                 如果有这个人名，则出现次数加 1
                    name_freq[flag][n] += 1
        f.close()
        print(name_freq)
        for news,freq in name_freq.items():
            print('news {0}----'.format(news))
            for n,f in freq.items():
                print('{0}:\t\t{1}'.format(n,f))
```

```
{1: {'Xuesen Qian': 4, 'Yonghuai Guo': 6, 'Jiaxian Deng': 1},
2: {'Ole Gunnar Solskjaer': 2, 'Massimiliano Allegri': 4, 'Emre Can': 1, 'Mario Mandzukic': 1},
3: {'Andrew Carnegie': 1, 'John D. Rockefeller': 1, 'Bill Gates': 1, 'Jeff Bezos': 3, 'Bob Iger': 1}}
news 1----
Xuesen Qian:                4
Yonghuai Guo:                6
Jiaxian Deng:                1
news 2----
Ole Gunnar Solskjaer:                2
Massimiliano Allegri:                4
Emre Can:                1
Mario Mandzukic:                1
news 3----
Andrew Carnegie:                1
John D. Rockefeller:                1
Bill Gates:                1
Jeff Bezos:                3
Bob Iger:                1
```

In [ ]: