需求：

- 读取文件usa_election.txt
- 查看文件样式及基本信息
- 新建一列各个候选人所在党派party
- 查看party这一列中有哪些元素
- 统计party列中各个元素出现次数
- 查看各个党派收到的政治献金总数contb_receipt_amt
- 查看每天各个党派收到的政治献金总数contb_receipt_amt
- 将表中日期格式转换为'yyyy-mm-dd'。日期格式,通过函数加map方式进行转换
- 得到每天各政党所收政治献金数目。
- 使用unstack()将上面所得数据中的party行索引变成列索引
- 查看职业为老兵DISABLED VETERAN的人主要支持谁，或者说查看老兵们捐赠给谁的钱最多
- 把索引变成列,Series变量.reset_index()
- 找出候选人的捐赠者中，捐赠金额最大的人的职业以及捐献额

In [14]:

```
# 月份和政党定义
months = {'JAN' : "01", 'FEB' : "02", 'MAR' : "03", 'APR' : "04", 'MAY' : "05", 'JUN
          'JUL' : "07", 'AUG' : "08", 'SEP' : "09", 'OCT': "10", 'NOV': "11", 'DEC'
of_interest = ['Obama, Barack', 'Romney, Mitt', 'Santorum, Rick',
               'Paul, Ron', 'Gingrich, Newt']
# 候选人党派映射
parties = {
  'Bachmann, Michelle': 'Republican',
  'Romney, Mitt': 'Republican',
  'Obama, Barack': 'Democrat',
  "Roemer, Charles E. 'Buddy' III": 'Reform',
  'Pawlenty, Timothy': 'Republican',
  'Johnson, Gary Earl': 'Libertarian',
  'Paul, Ron': 'Republican',
  'Santorum, Rick': 'Republican',
  'Cain, Herman': 'Republican',
  'Gingrich, Newt': 'Republican',
  'McCotter, Thaddeus G': 'Republican',
  'Huntsman, Jon': 'Republican',
  'Perry, Rick': 'Republican'
 }
```

In [15]:

```
import pandas as pd
import numpy as np
from pandas import DataFrame, Series
```

In [16]:

```
# 1. 读取文件usa_election.txt
df = pd.read_csv("./data/election/usa_election.txt")
```

```
/Users/guwanhua/venv36/lib/python3.6/site-packages/IPython/core/intera
ctiveshell.py:2785: DtypeWarning: Columns (6) have mixed types. Specif
y dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [17]:

```
# 2. 查看文件样式及基本信息
df.head(1)
```

Out[17]:

	cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_em
0	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RE

In [18]:

```
df.shape
```

Out[18]:

```
(536041, 16)
```

In [19]:

```
df.dtypes
```

Out[19]:

```
cmte_id                 object
cand_id                 object
cand_nm                 object
contbr_nm               object
contbr_city             object
contbr_st               object
contbr_zip              object
contbr_employer         object
contbr_occupation       object
contb_receipt_amt      float64
contb_receipt_dt        object
receipt_desc            object
memo_cd                 object
memo_text               object
form_tp                 object
file_num                 int64
dtype: object
```

In [20]:

```
# 3. 新建一列各个候选人所在党派party: 其中候选人字段是cand_nm, 这里使用map新建一列
df["party"] = df["cand_nm"].map(parties)
```

In [21]:

```
# 4. 查看party这一列中有哪些元素:  unique 去重
df["party"].unique()
```

Out[21]:

```
array(['Republican', 'Democrat', 'Reform', 'Libertarian'], dtype=objec
t)
```

In [22]:

```
# 5. 统计party列中各个元素出现次数：乍一看可能是需要对party中的元素分组进行统计个数，其实一个函数
```

value_counts()是Series中的，无参，返回一个带有每个元素出现次数的Series

In [23]:

```
df["party"].value_counts()
```

Out[23]:

```
Democrat       292400
Republican     237575
Reform           5364
Libertarian       702
Name: party, dtype: int64
```

In [24]:

```
# 6. 查看各个党派收到的政治献金总数contb_receipt_amt：分组聚合
df.groupby("party")["contb_receipt_amt"].sum()
```

Out[24]:

```
party
Democrat       8.105758e+07
Libertarian    4.132769e+05
Reform         3.390338e+05
Republican     1.192255e+08
Name: contb_receipt_amt, dtype: float64
```

In [25]:

```
# 7. 查看每天各个党派收到的政治献金总数contb_receipt_amt: 分组聚合，只不过这里的分组采用多个键
df.groupby(["contb_receipt_dt", "party"])["contb_receipt_amt"].sum()
```

Out[25]:

```
contb_receipt_dt  party
01-APR-11         Reform            50.00
                  Republican     12635.00
01-AUG-11         Democrat      175281.00
                  Libertarian     1000.00
                  Reform          1847.00
                  Republican    234598.46
01-DEC-11         Democrat      651532.82
                  Libertarian      725.00
                  Reform           875.00
                  Republican    486405.96
01-FEB-11         Republican       250.00
01-JAN-11         Republican      8600.00
01-JAN-12         Democrat       58098.80
                  Reform           515.00
                  Republican     75704.72
01-JUL-11         Democrat      165961.00
                  Libertarian     2000.00
                  Reform           100.00
                  Republican    115848.72
01-JUN-11         Democrat      145459.00
                  Libertarian      500.00
                  Reform            50.00
                  Republican    433109.20
01-MAR-11         Republican      1000.00
01-MAY-11         Democrat       82644.00
                  Reform           480.00
                  Republican     28663.87
01-NOV-11         Democrat      122529.87
                  Libertarian     3000.00
                  Reform          1792.00
                                  ...
30-OCT-11         Reform          3910.00
                  Republican     43913.16
30-SEP-11         Democrat     3373517.24
                  Libertarian      550.00
                  Reform          2050.00
                  Republican   4886331.76
31-AUG-11         Democrat      374387.44
                  Libertarian    10750.00
                  Reform           450.00
                  Republican   1017735.02
31-DEC-11         Democrat     3553072.57
                  Reform           695.00
                  Republican   1094376.72
31-JAN-11         Republican      6000.00
31-JAN-12         Democrat     1418410.31
                  Reform           150.00
                  Republican    869890.41
31-JUL-11         Democrat       20305.00
                  Reform           966.00
                  Republican     12781.02
31-MAR-11         Reform           200.00
                  Republican     62475.00
```

```
31-MAY-11          Democrat        351705.66
                   Libertarian        250.00
                   Reform             100.00
                   Republican      301339.80
31-OCT-11          Democrat        204996.87
                   Libertarian       4250.00
                   Reform            3105.00
                   Republican      734601.83
Name: contb_receipt_amt, Length: 1183, dtype: float64
```

In [26]:

```
# 8. 将表中日期格式转换为'yyyy-mm-dd': 这个是数据映射, 使用map + 自定义函数即可
def func(s):
    day, month, year = s.split("-")
    month = months[month]
    return "20%s-%s-%s"%(year, month, day)

df["contb_receipt_dt"] = df["contb_receipt_dt"].map(func)
```

In [27]:

```
df.head(1)
```

Out[27]:

	cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_em
0	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RE

In [31]:

```
# 9. 得到每天各政党所收政治献金数目： 多个键的分组
df.groupby(["contb_receipt_dt", "party"])["contb_receipt_amt"].sum()
```

Out[31]:

```
contb_receipt_dt  party      
2011-01-01        Republican       8600.00
2011-01-03        Republican       4800.00
2011-01-04        Republican       5000.00
2011-01-12        Republican       4150.00
2011-01-13        Republican       4000.00
2011-01-14        Republican       6000.00
2011-01-15        Republican        500.00
2011-01-16        Republican        750.00
2011-01-17        Republican        500.00
2011-01-18        Republican       4800.00
2011-01-20        Republican       2650.00
2011-01-21        Republican        250.00
2011-01-22        Republican        250.00
2011-01-24        Republican       2400.00
2011-01-26        Republican       5400.00
2011-01-27        Republican       2650.00
2011-01-28        Republican        650.00
2011-01-29        Republican        750.00
2011-01-31        Republican       6000.00
2011-02-01        Republican        250.00
2011-02-03        Republican       3250.00
2011-02-04        Republican       1000.00
2011-02-07        Republican       9300.00
2011-02-08        Republican       3000.00
2011-02-09        Republican       6550.00
2011-02-10        Republican        250.00
2011-02-11        Republican        250.00
2011-02-12        Republican        250.00
2011-02-13        Republican        250.00
2011-02-14        Republican       2500.00
                                     ...   
2012-01-22        Democrat        67194.23
                  Reform            450.00
                  Republican     507168.71
2012-01-23        Democrat       337307.07
                  Reform            225.00
                  Republican     645477.15
2012-01-24        Democrat       458909.23
                  Reform            500.00
                  Republican     462233.66
2012-01-25        Democrat       438949.32
                  Reform            282.00
                  Republican     416931.39
2012-01-26        Democrat       450268.94
                  Reform             25.00
                  Republican     256406.86
2012-01-27        Democrat       305785.47
                  Reform           3176.37
                  Republican     368441.82
2012-01-28        Democrat       235492.85
                  Reform            175.00
                  Republican      82775.80
2012-01-29        Democrat        93177.00
```

```
                  Reform              200.00
                  Republican        75220.02
2012-01-30        Democrat         435921.72
                  Reform              130.00
                  Republican       255204.80
2012-01-31        Democrat        1418410.31
                  Reform              150.00
                  Republican       869890.41
Name: contb_receipt_amt, Length: 1183, dtype: float64
```

使用unstack()将上面所得数据中的party行索引变成列索引, unstack()方法在groupby多个键的时候挺有用的

In [32]:

```
r1 = df.groupby(["contb_receipt_dt", "party"])["contb_receipt_amt"].sum()
r1.unstack().fillna(value=0)
```

Out[32]:

party	Democrat	Libertarian	Reform	Republican
contb_receipt_dt				
2011-01-01	0.00	0.0	0.00	8600.00
2011-01-03	0.00	0.0	0.00	4800.00
2011-01-04	0.00	0.0	0.00	5000.00
2011-01-12	0.00	0.0	0.00	4150.00
2011-01-13	0.00	0.0	0.00	4000.00
2011-01-14	0.00	0.0	0.00	6000.00
2011-01-15	0.00	0.0	0.00	500.00
2011-01-16	0.00	0.0	0.00	750.00
2011-01-17	0.00	0.0	0.00	500.00
2011-01-18	0.00	0.0	0.00	4800.00
2011-01-20	0.00	0.0	0.00	2650.00
2011-01-21	0.00	0.0	0.00	250.00
2011-01-22	0.00	0.0	0.00	250.00
2011-01-24	0.00	0.0	0.00	2400.00
2011-01-26	0.00	0.0	0.00	5400.00
2011-01-27	0.00	0.0	0.00	2650.00
2011-01-28	0.00	0.0	0.00	650.00
2011-01-29	0.00	0.0	0.00	750.00
2011-01-31	0.00	0.0	0.00	6000.00
2011-02-01	0.00	0.0	0.00	250.00
2011-02-03	0.00	0.0	0.00	3250.00
2011-02-04	0.00	0.0	0.00	1000.00
2011-02-07	0.00	0.0	0.00	9300.00
2011-02-08	0.00	0.0	0.00	3000.00
2011-02-09	0.00	0.0	0.00	6550.00
2011-02-10	0.00	0.0	0.00	250.00
2011-02-11	0.00	0.0	0.00	250.00
2011-02-12	0.00	0.0	0.00	250.00
2011-02-13	0.00	0.0	0.00	250.00
2011-02-14	0.00	0.0	0.00	2500.00
...	...	...	...	...
2012-01-02	89743.60	0.0	2437.13	114037.13

party	Democrat	Libertarian	Reform	Republican
contb_receipt_dt				
2012-01-03	87406.97	0.0	4006.32	155803.62
2012-01-04	166547.24	0.0	3445.80	577733.61
2012-01-05	198224.86	0.0	3925.48	451065.98
2012-01-06	138822.95	0.0	12676.24	262798.46
2012-01-07	91161.12	0.0	4201.12	148145.58
2012-01-08	81758.00	0.0	3457.52	84342.84
2012-01-09	206996.99	0.0	1950.00	501931.44
2012-01-10	191988.12	0.0	2195.00	487901.67
2012-01-11	185823.52	0.0	945.00	452916.99
2012-01-12	467212.53	0.0	625.00	348327.39
2012-01-13	374570.48	0.0	351.00	463368.26
2012-01-14	81687.80	0.0	200.00	608470.68
2012-01-15	72983.50	0.0	400.00	322194.08
2012-01-16	117163.21	0.0	400.00	367791.70
2012-01-17	298246.61	0.0	40.00	625365.77
2012-01-18	219002.47	0.0	0.00	888681.17
2012-01-19	275532.88	0.0	65.00	1066250.23
2012-01-20	245166.57	0.0	386.00	401298.03
2012-01-21	18513.50	0.0	280.00	374261.81
2012-01-22	67194.23	0.0	450.00	507168.71
2012-01-23	337307.07	0.0	225.00	645477.15
2012-01-24	458909.23	0.0	500.00	462233.66
2012-01-25	438949.32	0.0	282.00	416931.39
2012-01-26	450268.94	0.0	25.00	256406.86
2012-01-27	305785.47	0.0	3176.37	368441.82
2012-01-28	235492.85	0.0	175.00	82775.80
2012-01-29	93177.00	0.0	200.00	75220.02
2012-01-30	435921.72	0.0	130.00	255204.80
2012-01-31	1418410.31	0.0	150.00	869890.41

376 rows × 4 columns

In [33]:

```
# 10. 查看职业为老兵DISABLED VETERAN的人主要支持谁，或者说查看老兵们捐赠给谁的钱最多
df['contbr_occupation'] == 'DISABLED VETERAN'
```

Out[33]:

```
0         False
1         False
2         False
3         False
4         False
5         False
6         False
7         False
8         False
9         False
10        False
11        False
12        False
13        False
14        False
15        False
16        False
17        False
18        False
19        False
20        False
21        False
22        False
23        False
24        False
25        False
26        False
27        False
28        False
29        False
          ...
536011    False
536012    False
536013    False
536014    False
536015    False
536016    False
536017    False
536018    False
536019    False
536020    False
536021    False
536022    False
536023    False
536024    False
536025    False
536026    False
536027    False
536028    False
536029    False
536030    False
536031    False
536032    False
536033    False
```

```
536034    False
536035    False
536036    False
536037    False
536038    False
536039    False
536040    False
Name: contbr_occupation, Length: 536041, dtype: bool
```

In [36]:

```
r2 = df.loc[df['contbr_occupation'] == 'DISABLED VETERAN']
```

In [43]:

```
r2.groupby("cand_nm")['contb_receipt_amt'].sum().sort_values(ascending=False).index[
```

Out[43]:

```
'Obama, Barack'
```

补充一个reset_index的方法, 这个方法对于Series有时候是挺有用的

In [44]:

```
r2.groupby("cand_nm")['contb_receipt_amt'].sum()
```

Out[44]:

```
cand_nm
Cain, Herman        300.00
Obama, Barack      4205.00
Paul, Ron          2425.49
Santorum, Rick      250.00
Name: contb_receipt_amt, dtype: float64
```

In [45]:

```
# 11. 把索引变成列,Series变量.reset_index()
# 把cand_nm和这个Series的name都变为列索引
r2.groupby("cand_nm")['contb_receipt_amt'].sum().reset_index()
```

Out[45]:

	cand_nm	contb_receipt_amt
0	Cain, Herman	300.00
1	Obama, Barack	4205.00
2	Paul, Ron	2425.49
3	Santorum, Rick	250.00

In [49]:

```
# 12. 找出候选人的捐赠者中，捐赠金额最大的人的职业以及捐献额：因为涉及到条件的判断，这里可以考虑使
max_amt = df['contb_receipt_amt'].max()
```

In [55]:

```
df.query("contb_receipt_amt == @max_amt")[['contbr_occupation', "contb_receipt_amt"]
```

Out[55]:

	contbr_occupation	contb_receipt_amt
176127	NaN	1944042.43

In [ ]:

```

```