# PyCon 2015 大会 上海 - Python大数据分析与可视化

丁来强

- Email: wjo1212 at 163.com
- Wechat: LaiQiangDing


# CSDN 500W 客户账户分析

In [1]:

```
%matplotlib inline
%run ./env.py
```

# combine together

In [2]:

```
!ls *csd*
```

```
www.csdn.net.sql
```

In [3]:

```
!head -n 5 "www.csdn.net.sql"
```

```
zdg # 12344321 # zdg@csdn.net
LaoZheng # 670203313747 # chengming_zheng@163.com
fstao # 730413 # fstao@tom.com
huwolf # 2535263 # hujiye@263.net
cadcjl # KIC43dk6! # ccedcjl@21cn.com
```

In [5]:

```
# 读取
df = pd.read_csv("./www.csdn.net.sql",
                 header=None,
                 names=['name', 'password', 'email'],
                 sep=" # ")
```

In [6]:

```
df.head()
```

Out[6]:

| | name | password | email |
|---|---|---|---|
| **0** | zdg | 12344321 | zdg@csdn.net |
| **1** | LaoZheng | 670203313747 | chengming_zheng@163.com |
| **2** | fstao | 730413 | fstao@tom.com |
| **3** | huwolf | 2535263 | hujiye@263.net |
| **4** | cadcjl | KIC43dk6! | ccedcjl@21cn.com |

In [8]:

```
# 提取邮箱和邮箱名
df = pd.concat([df,
            df.email.str.extract(r"(?P<email_name>.+)@(?P<email_domain>.+)")],
                axis=1
          ).drop('email', axis=1)
```

In [9]:

```
# 小写邮箱
df = df.assign(email_domain=df.email_domain.str.lower())
```

In [10]:

```
# drop空邮箱
df = df.dropna()
```

In [11]:

```
df[:10]
```

Out[11]:

| | name | password | email_name | email_domain |
|---|---|---|---|---|
| **0** | zdg | 12344321 | zdg | csdn.net |
| **1** | LaoZheng | 670203313747 | chengming_zheng | 163.com |
| **2** | fstao | 730413 | fstao | tom.com |
| **3** | huwolf | 2535263 | hujiye | 263.net |
| **4** | cadcjl | KIC43dk6! | ccedcjl | 21cn.com |
| **5** | netsky | s12345 | songmail | 21cn.com |
| **6** | Michael | apple | appollp | netease.com |
| **7** | siclj | lj7202 | junlu | peoplemail.com.cn |
| **8** | jinbuhuan | 12345 | jinbuhuan | 163.net |
| **9** | Eie | hebeibdh | fwg | jxfw.com |

In [12]:

```
# 总数
df.count()
```

Out[12]:

```
name            6428604
password        6428604
email_name      6428604
email_domain    6428604
dtype: int64
```

In [13]:

```
# Top 邮箱
top_emails = df.email_domain.value_counts()
```

In [14]:

```
# 增加百分比
top_emails = top_emails.to_frame().assign(
    percent=top_emails/df.name.count()*100)
```

In [15]:

```
# 修改列名
top_emails = top_emails.reset_index()
top_emails.columns=['Domain', 'Count', 'Percent']
```

In [16]:

```
# Top Emails
top_emails[:20]
```

Out[16]:

| | Domain | Count | Percent |
|---|---|---|---|
| **0** | qq.com | 1976196 | 30.740671 |
| **1** | 163.com | 1766927 | 27.485392 |
| **2** | 126.com | 807895 | 12.567192 |
| **3** | sina.com | 351594 | 5.469212 |
| **4** | yahoo.com.cn | 205491 | 3.196510 |
| **5** | hotmail.com | 202948 | 3.156953 |
| **6** | gmail.com | 186843 | 2.906432 |
| **7** | sohu.com | 104735 | 1.629203 |
| **8** | yahoo.cn | 87048 | 1.354073 |
| **9** | tom.com | 72365 | 1.125672 |
| **10** | yeah.net | 53295 | 0.829029 |
| **11** | 21cn.com | 50709 | 0.788803 |
| **12** | vip.qq.com | 35119 | 0.546293 |
| **13** | 139.com | 29207 | 0.454329 |
| **14** | 263.net | 24779 | 0.385449 |
| **15** | sina.com.cn | 19156 | 0.297981 |
| **16** | live.cn | 18920 | 0.294310 |
| **17** | sina.cn | 18601 | 0.289347 |
| **18** | yahoo.com | 18454 | 0.287061 |
| **19** | foxmail.com | 16432 | 0.255608 |

In [17]:

```
top_emails[:20].Percent.sum()
```

Out[17]:

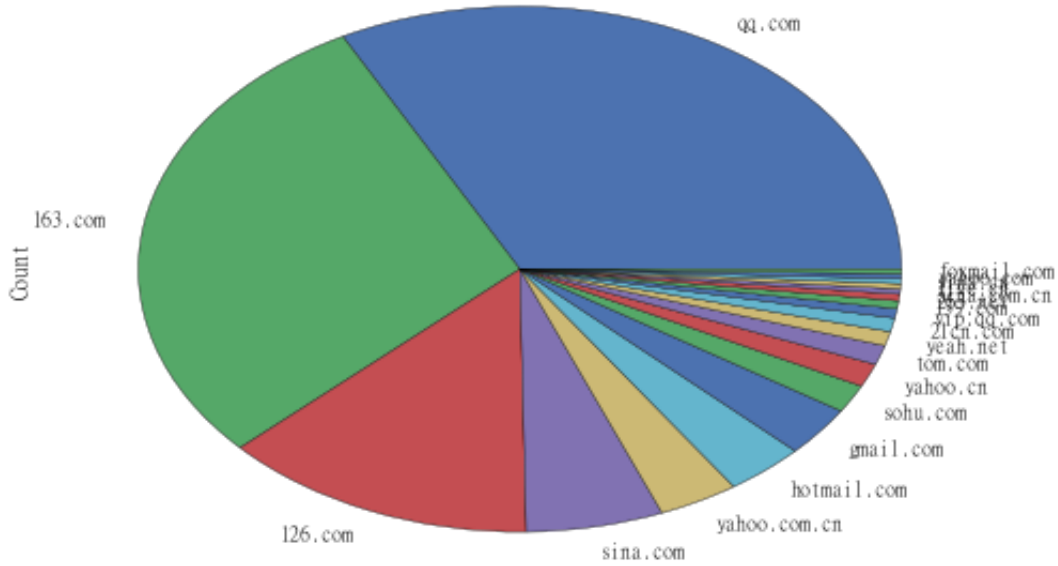
```
94.05951898732603
```

In [18]:

```
top_emails[:20].set_index('Domain').Count.plot(kind='pie')
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x13c0f6cd0>
```

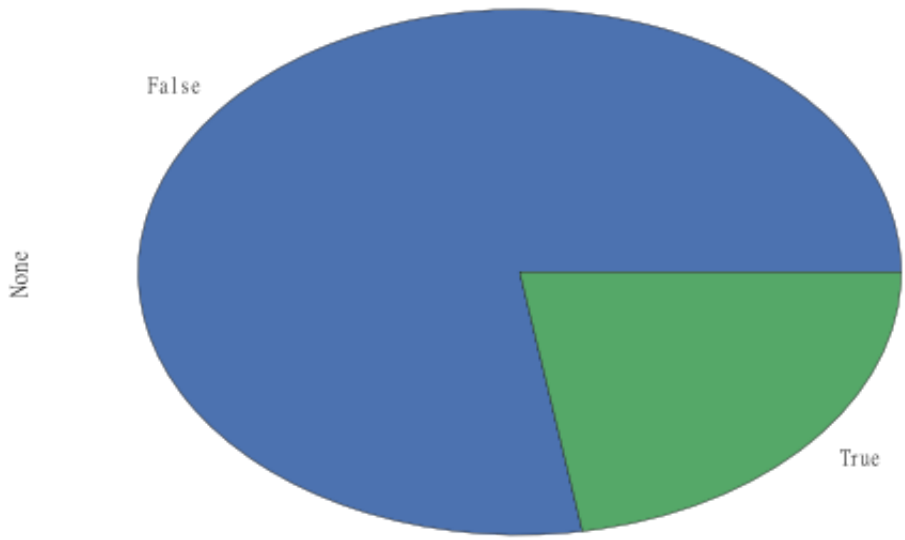



In [19]:

```
# 多少人用email名字注册了csdn
(df.name == df.email_name).value_counts().plot(kind='pie')
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1121b6310>
```



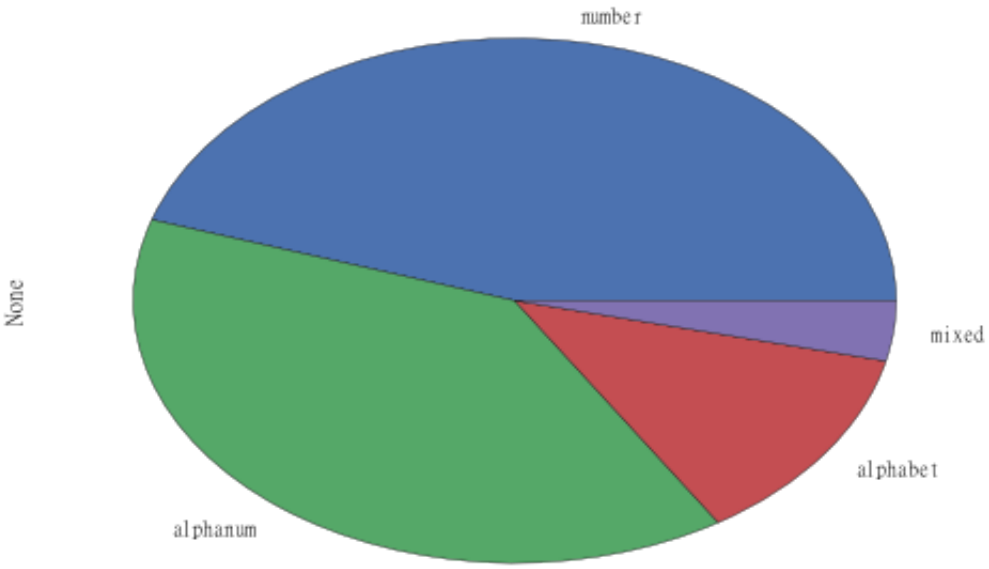
In [20]:

```
import re
def fun(x):
    if re.match(r'^\d+$', x):
        return 'number'
    if re.match(r'^[a-zA-Z]+$', x):
        return 'alphabet'
    if re.match(r'^[\da-zA-Z]+$', x):
        return 'alphanum'
    return 'mixed'

df.password.map(fun).value_counts().plot(kind='pie')
```

Out[20]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e7401410>
```




In [ ]: