数据爬取

来源：高德地图 http://map.amap.com/subway/index.html?&1100

In [1]:

```
import requests
import csv
import json
import time
from pyquery import PyQuery as pq
```

In [2]:

```
headers = {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_4) AppleWebKit/537.
}
```

In [3]:

```
#建立用于存放爬取数据的csv文件，创建好标题

def create_output_file():
    with open('./data/raw_data.csv', 'w') as f:
        csv_write = csv.writer(f)
        header = ['城市','线路','站名']
        csv_write.writerow(header)

create_output_file()
```

In [4]:

```
# 向此函数中传入城市相关信息，爬取地铁线路和站点数据，  并写入csv文件

def get_subway_info(id_info,citycode,cityname):
    url ='http://map.amap.com/service/subway?_1564387961003&srhdata={}_drw_{}.json'
    response = requests.get(url,headers)

    if response.status_code == 200:
        data = response.content.decode('utf-8')
        parsed = json.loads(data)

        for i in parsed['l']:
            title = i['kn']
            for item in i['st']:
                stops = item['n']
                with open('./data/raw_data.csv', 'a+') as f:
                    csv_write = csv.writer(f)
                    data_row = [cityname,title,stops]
                    csv_write.writerow(data_row)
```

In [5]:

```
# 本函数用于从高德地图网页中爬取有地铁的城市信息， 并将返回的信息传入 get_subway_info(id_info,

def get_city_info():

    url = 'http://map.amap.com/subway/index.html?&1100'
    response = requests.get(url,headers)

    if response.status_code == 200:
        data = pq(response.content.decode('utf-8'))

        for i in data('.city-list a.city').items():
            id_info  = i.attr('id')
            citycode = i.attr('cityname')
            cityname = i.text()
            get_subway_info(id_info,citycode,cityname)
            time.sleep(3)


        for i in data('.more-city-list a.other-city').items():
            id_info  = i.attr('id')
            citycode = i.attr('cityname')
            cityname = i.text()
            get_subway_info(id_info,citycode,cityname)
            time.sleep(3)
    else:
        print('Loading errors while crawling city info from target website')
```

In [6]:

```
if __name__ == '__main__':
    get_city_info()
```