各城市地铁信息爬虫与分析

综述

此项目致力于查看中国各城市的地铁分布与地铁站名字特点的情况.

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

此项目中,包括了数据的获取,解析,保存,评估,清洗,以及可视化.

我们利用pyquery,json,requests库从高德地图网站上爬取数据并存入本地csv文件,

之后用python加载文件,查看,评估,清洗数据,

最后通过pyecharts对数据进行探索性分析与可视化.

文件描述

- crawler.ipynb： 数据的爬虫解析保存过程
- analysis.ipynb：数据的评估，清洗，可视化过程
- raw_data.csv： 通过爬虫抓取的原始数据
- README.md

探索性数据分析

完成数据获取，解析，清洗后，我们将会去探讨以下几方面的问题：

都有哪些城市拥有地铁系统?

全国的地铁线路和站点数量有多少?

各个城市的地铁线路和站点数量是怎样分布的?

哪个城市哪条线路上的站点数量最多?

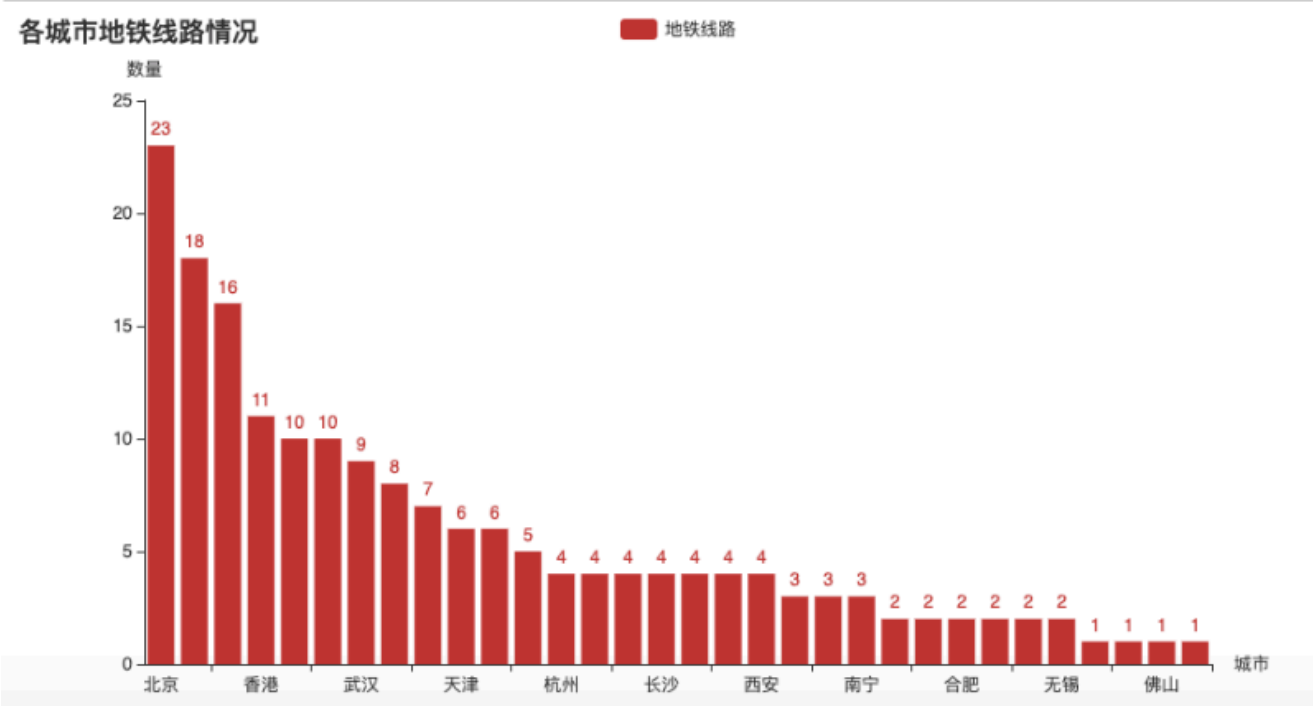
地铁站名字的词云图是怎样的?

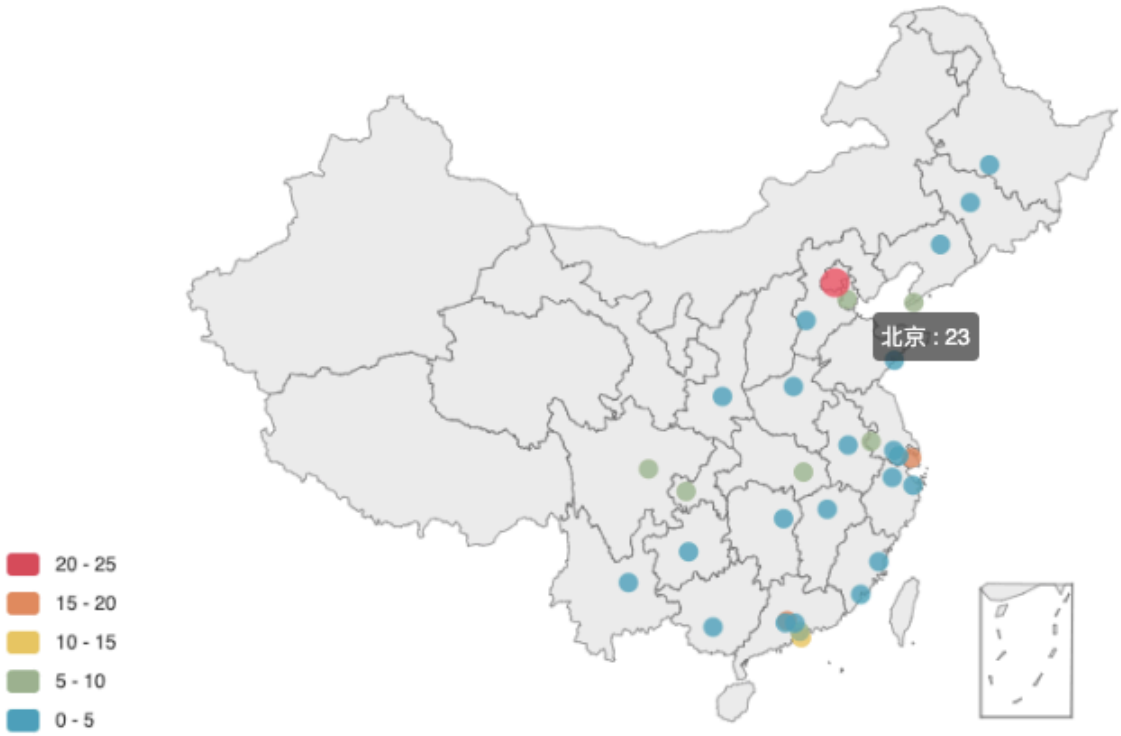
地铁站名的高频字有哪些?

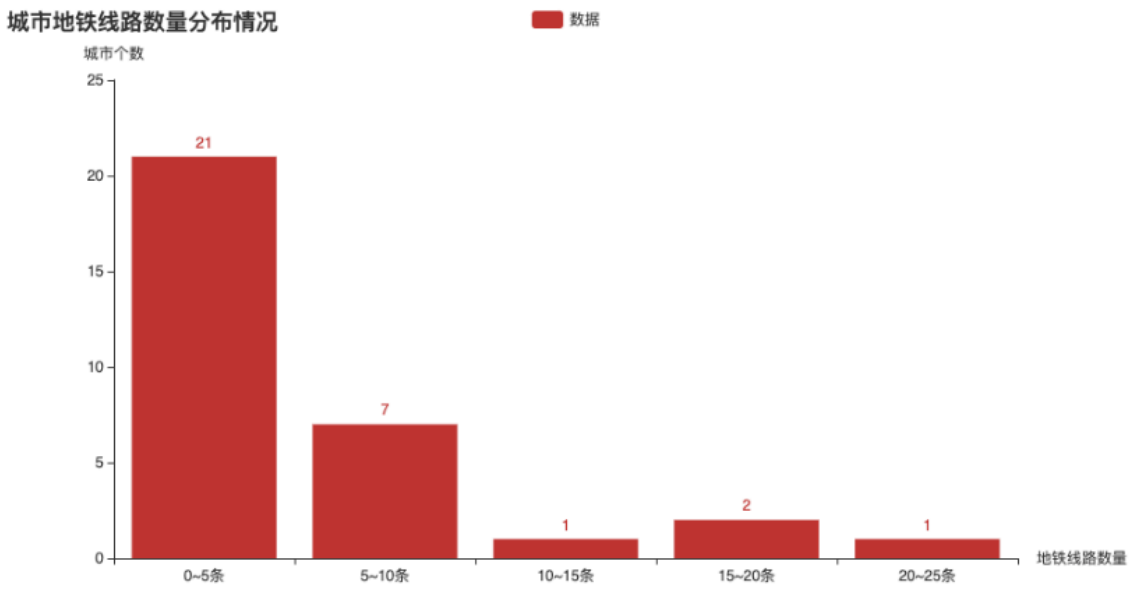
可视化报告

通过pandas统计,我们看到共有32个城市，182条地铁线路,3183个地铁站点

我们可以看到地铁大多分布在东部发达城市

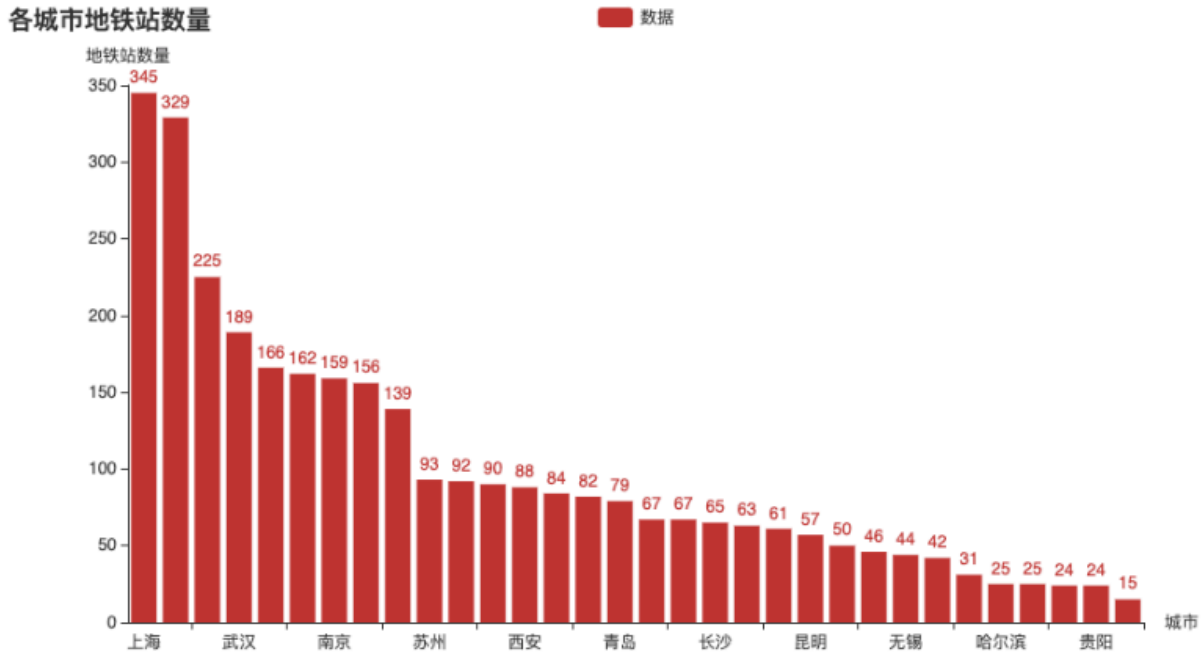






我们可以看到大部分城市所拥有的地铁线路在5条左右

上海,北京,广州,武汉,深圳所拥有的地铁站数量排在了前五名

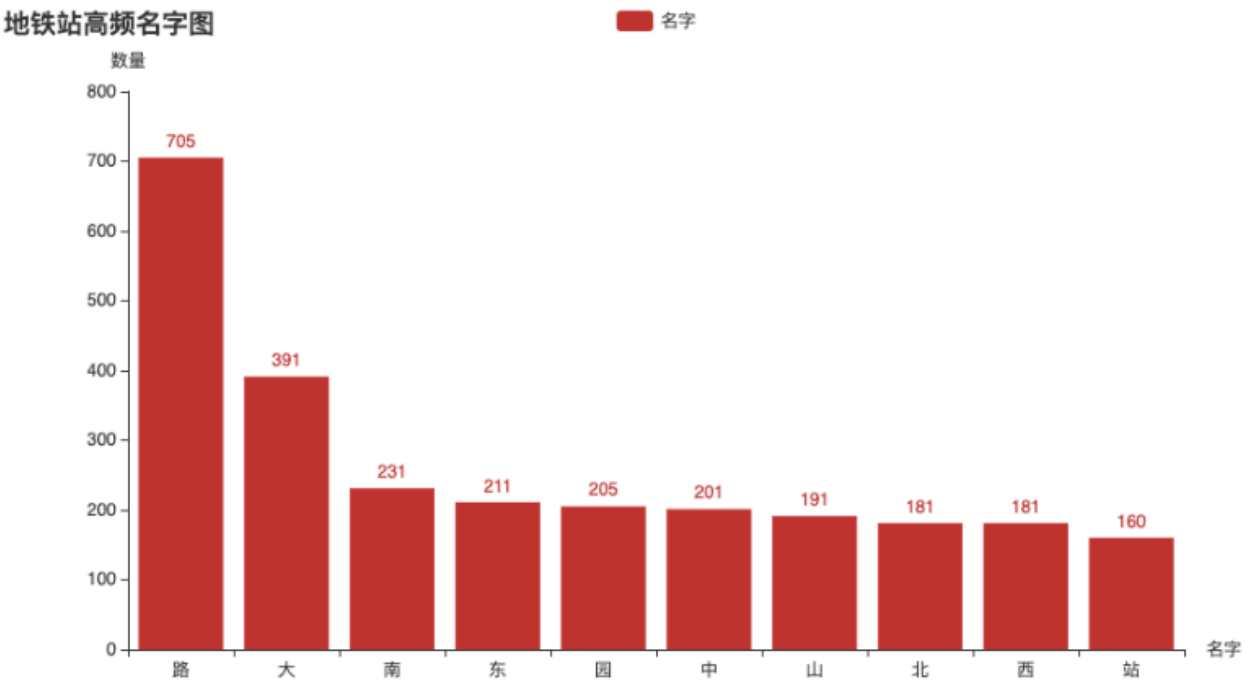


北京十号线的地铁站数量最多,达到了45个

```
# 哪个城市哪条线路地铁站数最多
data_lines.sort_values(by='站名',ascending=False)
```

	城市	线路	站名
20	北京	10号线	45
151	重庆	3号线	39
1	上海	11号线	38
74	天津	6号线	38
114	武汉	2号线	38
...	...	...	...
106	昆明	6号线	4
40	北京	机场线	4
166	长沙	磁浮快线	3
17	上海	磁悬浮	2
180	香港	迪士尼線	2

182 rows × 3 columns



我们可以看到’路’,’大’,’南’等字被用到的频率比较高

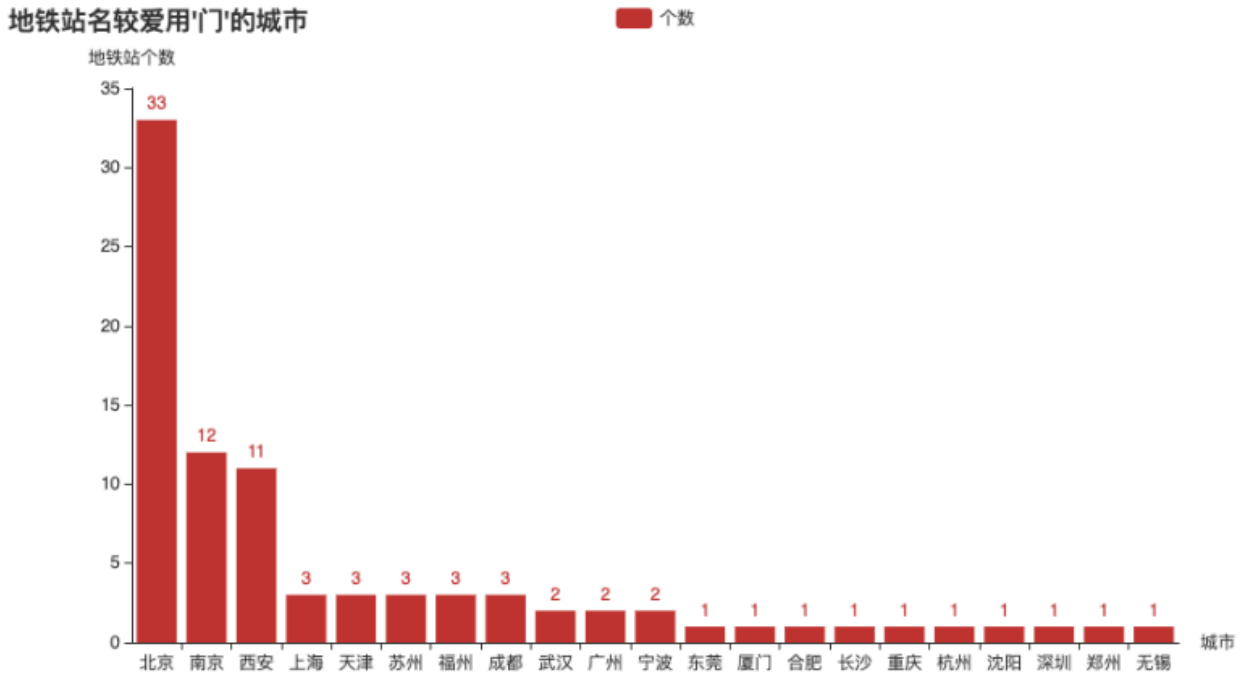
上海共计345个地铁站中,有210个站点都含有路字

```
: df_sh = df_stations[df_stations['城市']=='上海']
  len(df_sh[df_sh['站名'].str.contains('路')])

: 210
```

```
: df_stations.groupby(['城市'])['数量'].count()

: 城市
  上海     345
  东莞      15
  佛山      25
  北京     329
  南京     159
          ...
  重庆     162
  长春      84
  长沙      65
  青岛      79
  香港      92
  Name: 数量, Length: 32, dtype: int64
```



北京,南京,西安,这三个历史底蕴丰厚的城市都爱用’门’做为地铁站的名字