Data Analysis Lagou

Get data from lagou

In [1]:

```
import time
import pandas
import requests
```

get target position data

In [2]:

```
def getGoalData(data):
    for i in range(15):  # 每页默认15个职位
        info = {
            'positionName': data[i]['positionName'],  # 职位简称
            'companyShortName': data[i]['companyShortName'],  # 平台简称
            'salary': data[i]['salary'],  # 职位薪水
            'createTime': data[i]['createTime'],  # 发布时间
            'companyId': data[i]['companyId'],  # 公司ID
            'companyFullName': data[i]['companyFullName'],  # 公司全称
            'companyLabelList': data[i]['companyLabelList'],  # 公司规模
            'financeStage': data[i]['financeStage'],  # 融资情况
            'positionLables': data[i]['positionLables'],  # 所在行业
            'skillLables': data[i]['skillLables'],
            'education': data[i]['education'],  # 教育背景
            'district': data[i]['district'],  # 公司所在区域
            'workYear': data[i]['workYear']  # 区域详细地
        }
        data[i] = info
    return data
```

save data as csv file

In [3]:

```
def saveData(data):
    table = pandas.DataFrame(data)
#     table.to_csv('LaGou1.csv', index=False, mode='a+')
```

constant definition

In [4]:

```
header = {
    'Accept': 'application/json, text/javascript, */*; q=0.01',
    'Referer': 'https://www.lagou.com/jobs/list_%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHT
    'Host': 'www.lagou.com'}

url1 = 'https://www.lagou.com/jobs/list_%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98?labelWo
url = 'https://www.lagou.com/jobs/positionAjax.json?city=%E4%B8%8A%E6%B5%B7&needAddt
pages = 26
```

get and save data

In [5]:

```
for page in range(1, pages):
    form = {
        'first': 'false',
        'pn': page,
        'kd': '数据挖掘'
    }
    s = requests.Session()  # 建立session
    s.get(url=url1, headers=header, timeout=3)
    cookie = s.cookies  # 获取cookie
    respon = s.post(url=url, headers=header, data=form, cookies=cookie, timeout=3)
    time.sleep(8)
    result = respon.json()
    data = result['content']['positionResult']['result']  # 返回结果在preview中的具体返
    try:
        data_goal = getGoalData(data)
        saveData(data_goal)
    except IndexError:
        break
```

read the csv file and analyse

In [6]:

```
from jieba_fast import analyse
import pandas as pd
from pyecharts import Geo
from pyecharts import Pie
from pyecharts import WordCloud
from pyecharts import Funnel
from pyecharts import Bar
```

```
ERROR:lml.utils:failed to import pyecharts_snapshot
Traceback (most recent call last):
  File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.
6/site-packages/lml/utils.py", line 43, in do_import
    plugin_module = __import__(plugin_module_name)
ModuleNotFoundError: No module named 'pyecharts_snapshot'
```

In [7]:

```
data = pd.read_csv('LaGou1.csv')  # 读取数据
data.head()
```

Out[7]:

	positionName	companyShortName	salary	createTime	companyId	companyFullName	compa
0	数据挖掘	The NetCircle	18k-25k	2019-12-09 16:52:22	4670	人英网络（上海）有限公司	['年终 金', '3
1	数据挖掘工程师（2020校招）	莉莉丝游戏	10k-20k	2019-12-09 15:14:28	1938	上海莉莉丝科技股份有限公司	['都是 奖金',
2	数据挖掘	微创软件	30k-35k	2019-12-09 15:10:53	124652	上海微创软件股份有限公司	['绩效 假', '5
3	算法工程师	NextTao 互道信息	18k-30k	2019-12-09 17:10:55	56474	互道信息技术（上海）有限公司	['节日 训', '4
4	算法工程师	趣头条	25k-50k	2019-12-09 17:08:57	202104	上海基分文化传播有限公司	['专项 假', '5

data cleaning

In [8]:

```
# 去除实习岗位和地区为空的岗位
data = data[~data['positionName'].str.contains('intern|实习|产品')]
data = data[~data['district'].isnull()]
data = data[~data['district'].str.contains('district')]
data = data.reset_index(drop=True)
data.head()
```

Out[8]:

	positionName	companyShortName	salary	createTime	companyId	companyFullName	compa
0	数据挖掘	The NetCircle	18k-25k	2019-12-09 16:52:22	4670	人英网络（上海）有限公司	['年终金', ''
1	数据挖掘工程师（2020校招）	莉莉丝游戏	10k-20k	2019-12-09 15:14:28	1938	上海莉莉丝科技股份有限公司	['都是奖金',
2	数据挖掘	微创软件	30k-35k	2019-12-09 15:10:53	124652	上海微创软件股份有限公司	['绩效假', ''
3	算法工程师	NextTao 互道信息	18k-30k	2019-12-09 17:10:55	56474	互道信息技术（上海）有限公司	['节日训', ''
4	算法工程师	趣头条	25k-50k	2019-12-09 17:08:57	202104	上海基分文化传播有限公司	['专项假', ''

draw the heat map of job distribution in Shanghai and data mining

In [9]:

```
get_district = data.groupby(['district']).count()['positionName'].index.tolist()
count_district = data.groupby(['district']).count()['positionName'].tolist()
get_district = get_district + ['松江区', '奉贤区', '金山区']
count_district = count_district + ['0', '1', '1']
geo = Geo("上海地区‘数据挖掘‘职位分布热力图", "data from lagou", title_color="#fff", titl
          width=800, height=600,
          background_color='#404a59')
geo.add("职位分布热力图", get_district, count_district, visual_range=[0, 200], type='e
        visual_text_color="#fff", symbol_size=15,
        is_visualmap=True, is_roam=True, maptype="上海")  # type有scatter, effectScat
geo
```

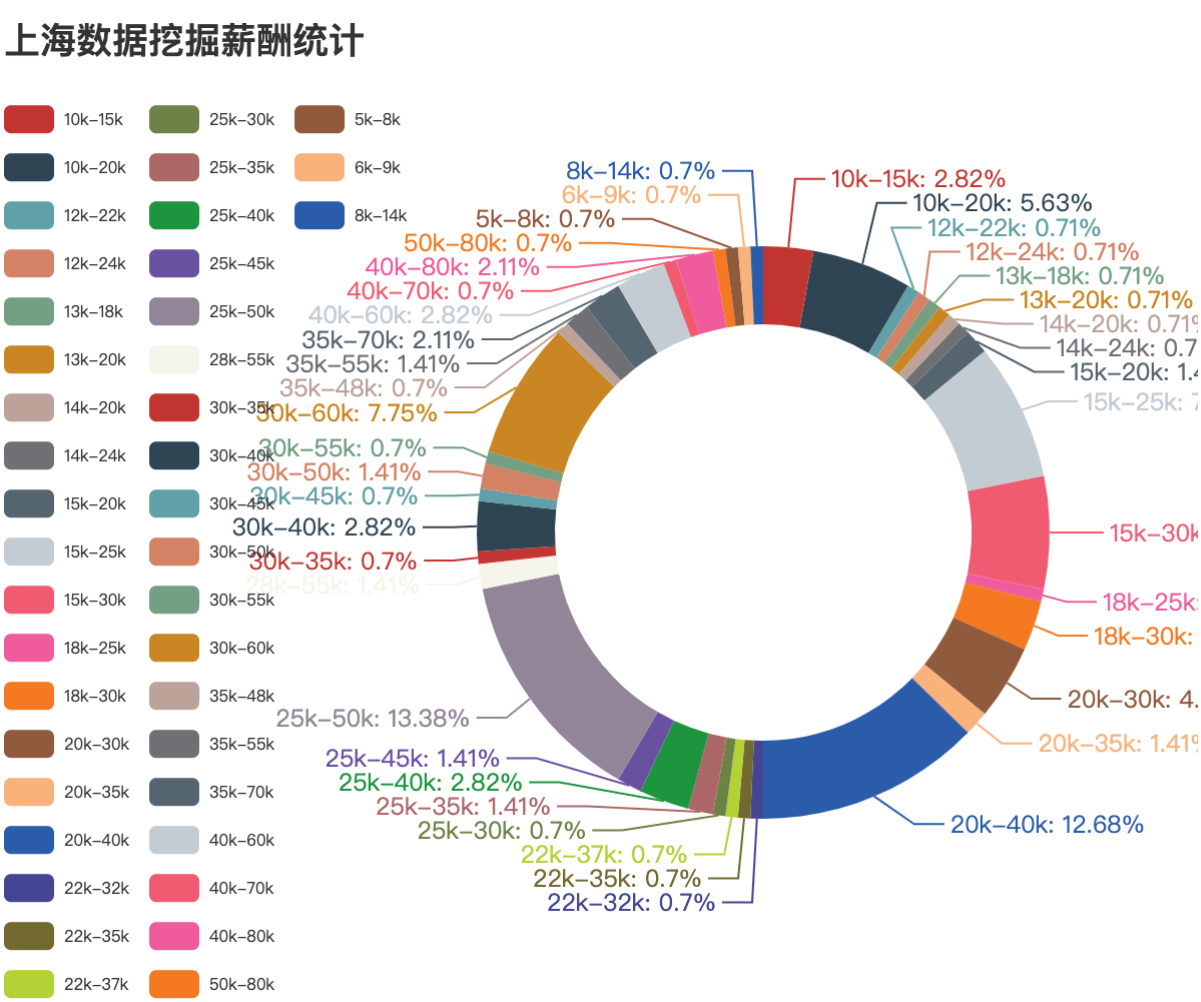
Out[9]:

Draw the data mining education pie chart in Shanghai

In [10]:

```
get_salary = data.groupby(['salary']).count()['positionName'].index.tolist()
count_salary = data.groupby(['salary']).count()['positionName'].tolist()
get_workYear = data.groupby(['workYear']).count()['positionName'].index.tolist()
count_workYear = data.groupby(['workYear']).count()['positionName'].tolist()
pie_salary = Pie("上海数据挖掘薪酬统计", title_pos='left',width=640, height=520)
pie_salary.add("", get_salary, count_salary, center=[60, 50], radius=[40, 55], label
        is_label_show=True, legend_orient='vertical',
        legend_pos='left',legend_text_size= 8, legend_top= "8%")
pie_workYear = Pie("上海地区数据挖掘工作年限统计", "data from lagou", title_pos='left',
pie_workYear.add("工资", get_workYear, count_workYear, center=[50, 50], is_legend_sh
# pie_workYear.render(path="上海地区'数据挖掘'工作年限饼状图.html")
pie_salary
```

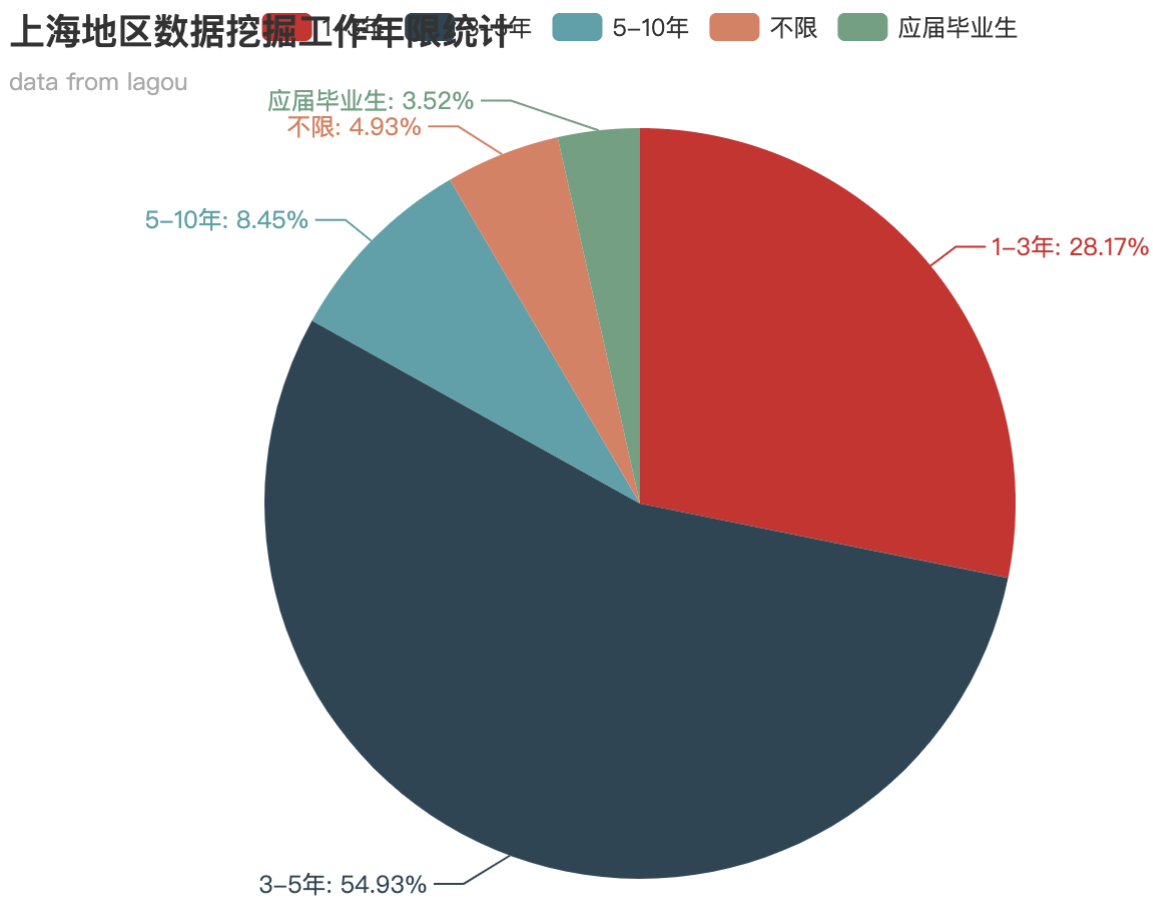
Out[10]:



In [11]:

```
pie_workYear
```

Out[11]:



Draw the "data mining" word cloud in Shanghai area

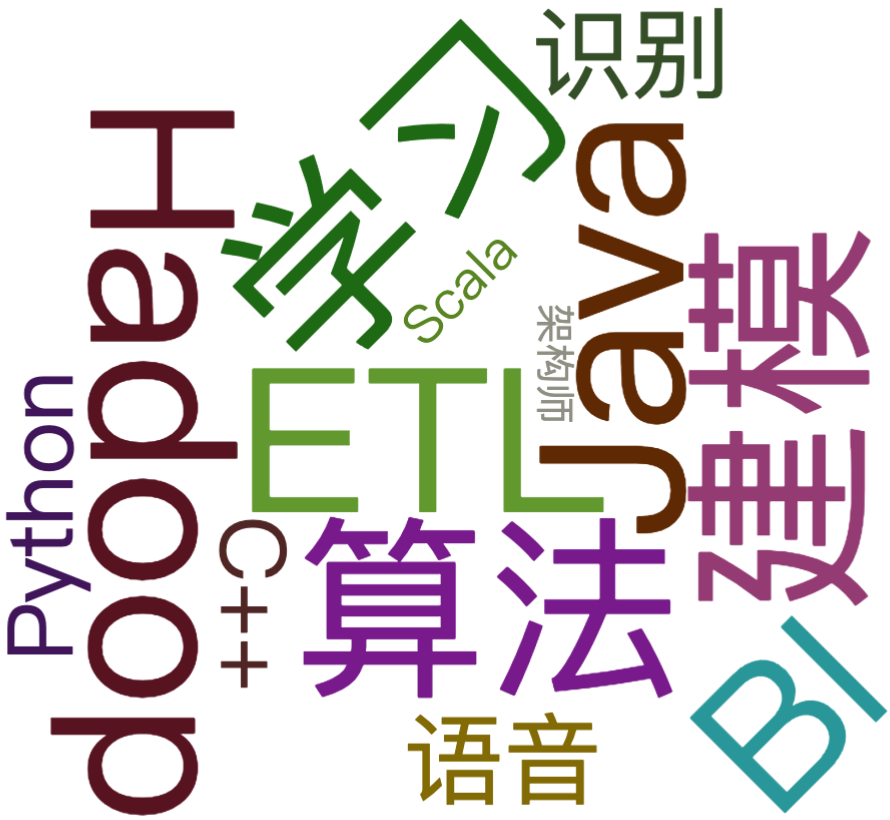
In [12]:

```
text = ''
counts = {}
for i in range(len(data['skillLables'])):
    content = data['skillLables'][i].strip()
    text += content
    tags = analyse.extract_tags(text, topK=100, withWeight=False)
    for tag in tags:  # 遍历方法统计词频
        if len(tag) == 1:
            continue
        else:
            counts[tag] = counts.get(tag, 0) + 1
count_skillLables = list(counts.values())
get_skillLables = list(counts.keys())
myWordCloud = WordCloud("绘制词云", width=680, height=520)
myWordCloud.add("", get_skillLables, count_skillLables, word_size_range=[20, 100])
myWordCloud
```

```
Building prefix dict from the default dictionary ...
DEBUG:jieba_fast:Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/xy/99lj18yj43qc9kttrs_b2v6c0000g
n/T/jieba.cache
DEBUG:jieba_fast:Loading model from cache /var/folders/xy/99lj18yj43qc
9kttrs_b2v6c0000gn/T/jieba.cache
Loading model cost 1.180 seconds.
DEBUG:jieba_fast:Loading model cost 1.180 seconds.
Prefix dict has been built succesfully.
DEBUG:jieba_fast:Prefix dict has been built succesfully.
```

Out[12]:

绘制词云



In [13]:

```
text = ''
counts = {}
for i in range(len(data['positionLables'])):
    content = data['positionLables'][i].strip()
    text += content
    tags = analyse.extract_tags(text, topK=100, withWeight=False)
    for tag in tags:  # 遍历方法统计词频
        if len(tag) == 1:
            continue
        else:
            counts[tag] = counts.get(tag, 0) + 1
count_skillLables = list(counts.values())
get_skillLables = list(counts.keys())
myWordCloud = WordCloud("数据挖掘标签", width=680, height=520)
myWordCloud.add("", get_skillLables, count_skillLables, word_size_range=[20, 100])
myWordCloud
```

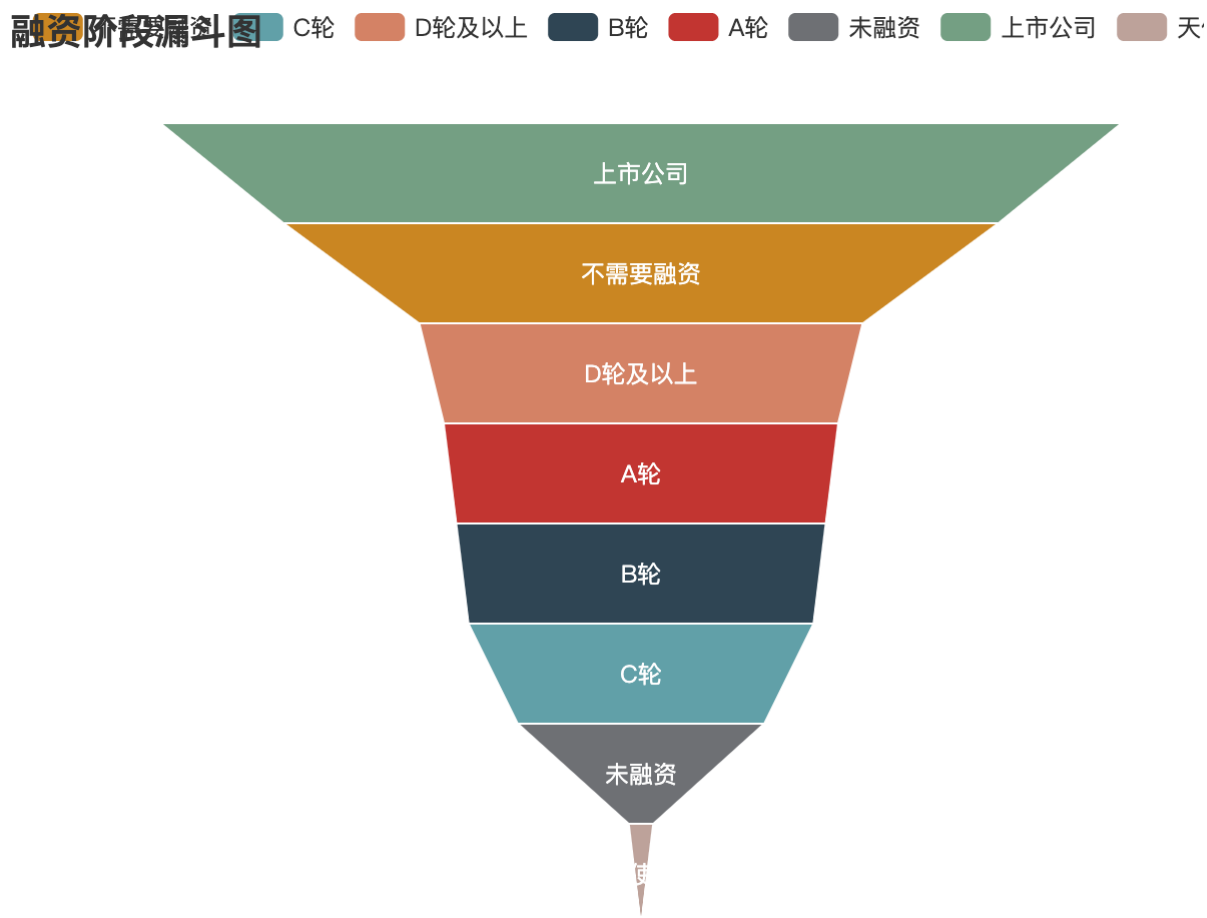
Out[13]:

数据挖掘标签

In [14]:

```
get_financeStage = data.groupby(['financeStage']).count()['positionName'].index.toli
count_financeStage = data.groupby(['financeStage']).count()['positionName'].tolist()
funnel = Funnel("融资阶段漏斗图", width=640, height=520)
funnel.add("融资阶段", get_financeStage, count_financeStage, is_label_show=True, labe
funnel
```

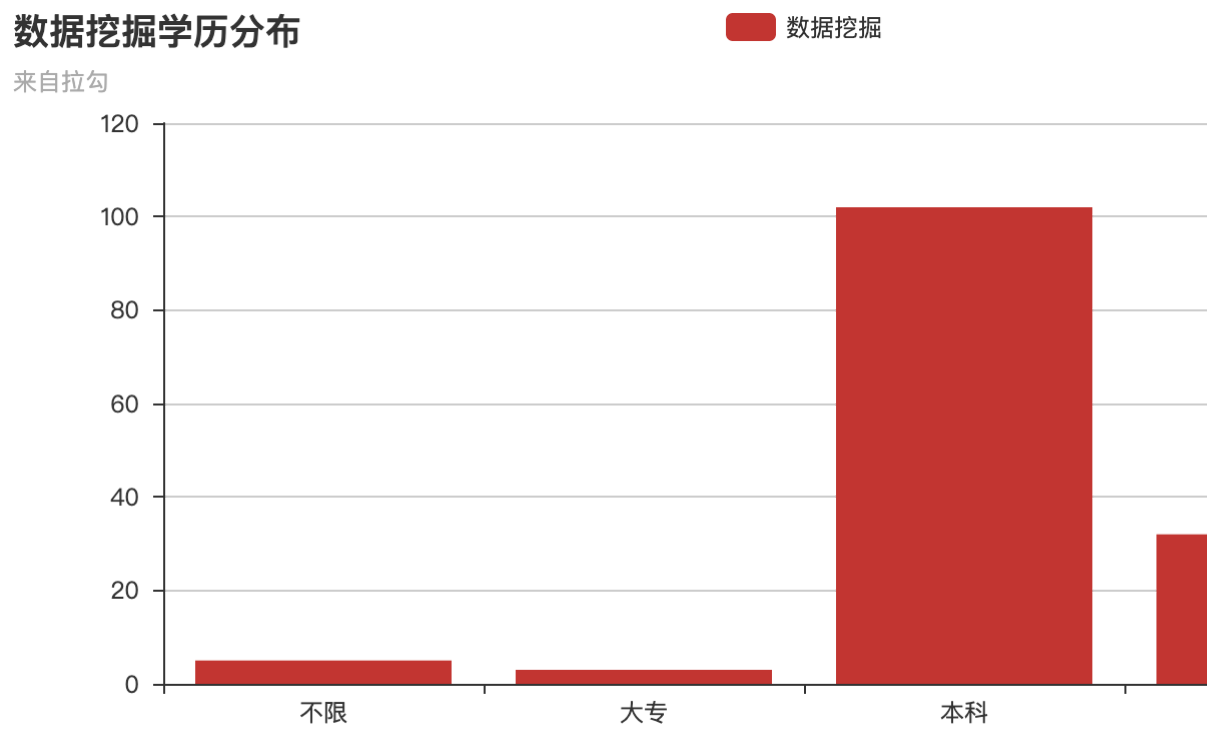
Out[14]:



In [15]:

```
get_education = data.groupby(['education']).count()['positionName'].index.tolist()
count_education = data.groupby(['education']).count()['positionName'].tolist()
bar = Bar("数据挖掘学历分布", "来自拉勾")
bar.add("数据挖掘", get_education, count_education)
bar
```

Out[15]:



In [ ]:

```

```