In [1]:

```
import requests
import pandas as pd
from lxml import etree

html = 'https://ncov.dxy.cn/ncovh5/view/pneumonia'
html_data = requests.get(html)
html_data.encoding = 'utf-8'
html_data = etree.HTML(html_data.text, etree.HTMLParser())
html_data = html_data.xpath(
    '//*[@id="getListByCountryTypeService2true"]/text()')  # xpath方法选择疫情的数据集合
ncov_world = html_data[0][49:-12]
ncov_world = ncov_world.replace('true', 'True')
ncov_world = ncov_world.replace('false', 'False')
ncov_world = eval(ncov_world)

country = []
confirmed = []
lived = []
dead = []

for i in ncov_world:  # 分离国家名称，确诊人数，治愈人数和死亡人数并存入dataframe里备用
    country.append(i['provinceName'])
    confirmed.append(i['confirmedCount'])
    lived.append(i['curedCount'])
    dead.append(i['deadCount'])

data_world = pd.DataFrame()
data_world['国家名称'] = country
data_world['确诊人数'] = confirmed
data_world['治愈人数'] = lived
data_world['死亡人数'] = dead
data_world.head(5)
```

Out[1]:

	国家名称	确诊人数	治愈人数	死亡人数
0	法国	29583616	368023	149044
1	德国	26200663	4328400	138781
2	韩国	18053287	336548	24103
3	英国	22455392	6491069	178880
4	西班牙	12311477	150376	106105

In [2]:

```
import pandas as pd
data_world = pd.read_csv('https://labfile.oss.aliyuncs.com/courses/2791/data_world.csv')
data_world.head(5)
```

Out[2]:

	国家名称	确诊人数	治愈人数	死亡人数
0	法国	27626578	368023	144130
1	德国	23376879	4328400	132929
2	韩国	16212751	336548	20889
3	英国	21819851	6491069	171560
4	西班牙	11662214	150376	103266

In [3]:

```
data_economy = pd.read_csv(
    "https://labfile.oss.aliyuncs.com/courses/2791/gpd_2016_2020.csv", index_col=0)
time_index = pd.date_range(start='2016', periods=18, freq='Q')
data_economy.index = time_index
data_economy
```

Out[3]:

	国内生产总值	第一产业增加值	第二产业增加值	第三产业增加值	农林牧渔业增加值	工业增加值	制造业增加值	建筑业增加值	批发和零售业增加值	交通运输、储存
2016-03-31	162410.0	8312.7	61106.8	92990.5	8665.5	53666.4	45784.0	7763.0	16847.5	718
2016-06-30	181408.2	12555.9	73416.5	95435.8	13045.5	60839.2	52378.3	12943.8	17679.8	829
2016-09-30	191010.6	17542.4	75400.5	98067.8	18162.2	61902.5	52468.3	13870.6	18513.0	859
2016-12-31	211566.2	21728.2	85504.1	104334.0	22577.8	68998.4	58878.4	16921.5	20684.1	896
2017-03-31	181867.7	8205.9	69315.5	104346.3	8595.8	60909.3	51419.7	8725.3	18608.9	809
2017-06-30	201950.3	12644.9	82323.0	106982.4	13204.2	68099.8	58172.1	14574.4	19473.6	939
2017-09-30	212789.3	18255.8	84574.1	109959.5	18944.2	69327.2	58632.6	15590.1	20342.9	968
2017-12-31	235428.7	22992.9	95368.0	117067.8	23915.8	76782.9	65652.1	19015.8	22731.1	994
2018-03-31	202035.7	8575.7	76598.2	116861.8	9005.8	66905.6	56631.9	10073.8	20485.5	880
2018-06-30	223962.2	13003.8	91100.6	119857.8	13662.2	75122.1	64294.9	16404.3	21374.2	1017
2018-09-30	234474.3	18226.9	93112.5	123134.9	18961.8	76239.6	64348.2	17294.5	22334.1	1058
2018-12-31	258808.9	24938.7	104023.9	129846.2	25929.0	82822.1	70662.1	21720.4	24710.0	1077
2019-03-31	218062.8	8769.4	81806.5	127486.9	9249.4	71064.5	60357.1	11143.1	21959.2	938
2019-06-30	242573.8	14437.6	97315.6	130820.6	15108.7	79820.7	68041.8	17954.2	23097.0	1086
2019-09-30	252208.7	19798.0	97790.4	134620.4	20629.0	79501.8	66823.8	18734.6	23993.6	1131
2019-12-31	278019.7	27461.6	109252.8	141305.2	28579.9	86721.6	73952.4	23072.4	26795.9	1124
2020-03-31	206504.3	10186.2	73638.0	122680.1	10708.4	64642.0	53852.0	9377.8	18749.6	786
2020-06-30	250110.1	15866.8	99120.9	135122.3	16596.4	80402.4	69258.8	19156.8	23696.1	1065

In [4]:

```
data_area = pd.read_csv('https://labfile.oss.aliyuncs.com/courses/2791/DXYArea.csv')
data_news = pd.read_csv('https://labfile.oss.aliyuncs.com/courses/2791/DXYNews.csv')
```

In [5]:

```
data_area = data_area.loc[data_area['countryName'] == data_area['provinceName']]
data_area_times = data_area[['countryName', 'province_confirmedCount',
                             'province_curedCount', 'province_deadCount', 'updateTime']]

time = pd.DatetimeIndex(data_area_times['updateTime'])  # 根据疫情的更新时间来生成时间序列
data_area_times.index = time  # 生成索引
data_area_times = data_area_times.drop('updateTime', axis=1)
data_area_times.head(5)

data_area_times.isnull().any()  # 查询是否有空值
```

Out[5]:

```
countryName               False
province_confirmedCount   False
province_curedCount       False
province_deadCount        False
dtype: bool
```

In [6]:

```
data_news_times = data_news[['pubDate', 'title', 'summary']]
time = pd.DatetimeIndex(data_news_times['pubDate'])
data_news_times.index = time  # 生成新闻数据的时间索引
data_news_times = data_news_times.drop('pubDate', axis=1)
data_news_times.head(5)
```

Out[6]:

pubDate	title	summary
2020-07-17 05:40:08	美国新增71434例新冠肺炎确诊病例，累计确诊超354万例	据美国约翰斯·霍普金斯大学统计数据显示，截至美东时间7月16日17:33时（北京时间17日0...
2020-07-17 06:06:49	巴西新冠肺炎确诊病例破201万，近六成大城市确诊病例加速增长	截至当地时间7月16日18时，巴西新增新冠肺炎确诊病例45403例，累计确诊2012151例...
2020-07-16 22:31:00	阿塞拜疆新增493例新冠肺炎确诊病例 累计确诊26165例	当地时间7月16日，阿塞拜疆国家疫情防控指挥部发布消息，在过去24小时内，阿塞拜疆新增新冠肺...
2020-07-16 22:29:48	科威特新增791例新冠肺炎确诊病例 累计确诊57668例	科威特卫生部当地时间16日下午发布通告，确认过去24小时境内新增791例新冠肺炎确诊病例，同...
2020-07-16 21:26:54	罗马尼亚新增777例新冠肺炎确诊病例 累计确诊35003例	据罗马尼亚政府7月16日公布的数据，过去24小时对19097人进行新冠病毒检测，确诊777例...

In [7]:

```
print(data_world.isnull().any())
print(data_economy.isnull().any())
print(data_area_times.isnull().any())
print(data_news_times.isnull().any())  # 确认各个数据集是否空集
```

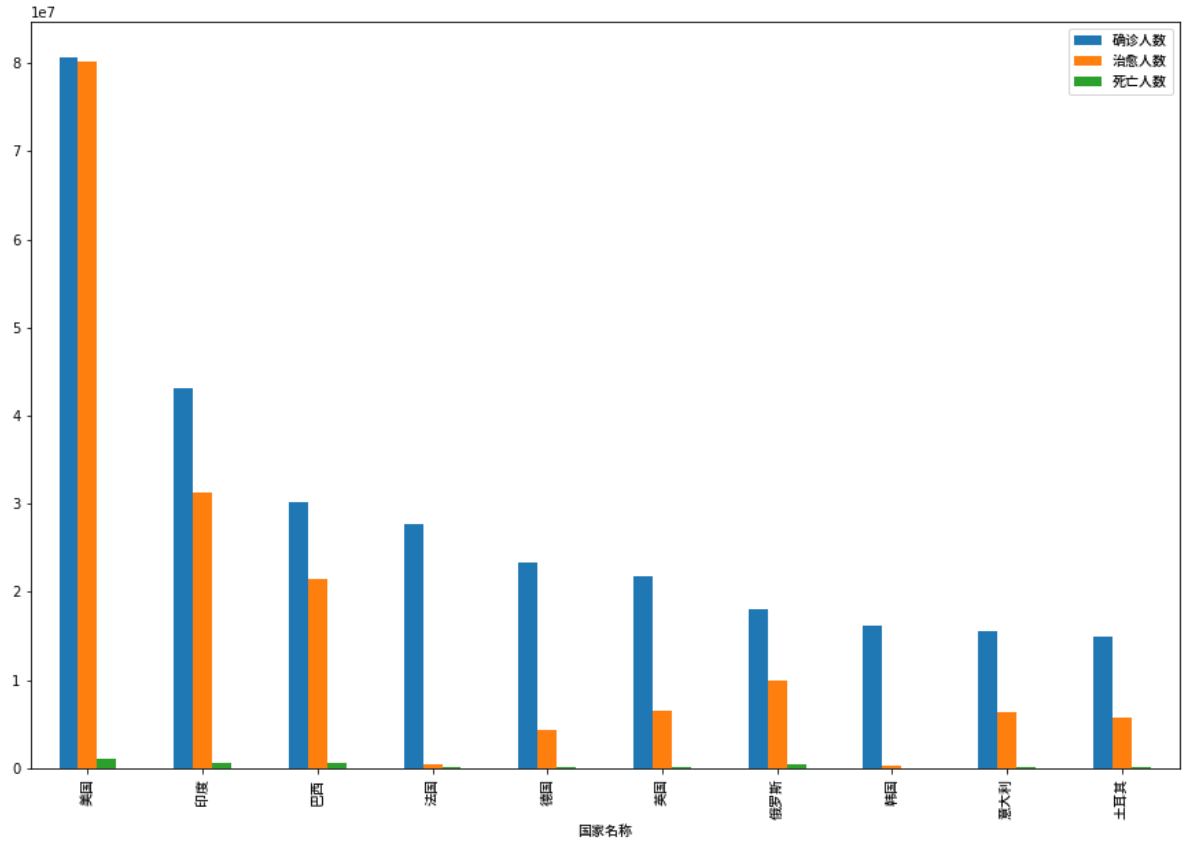
```
国家名称     False
确诊人数     False
治愈人数     False
死亡人数     False
dtype: bool
国内生产总值                 False
第一产业增加值                False
第二产业增加值                False
第三产业增加值                False
农林牧渔业增加值               False
工业增加值                  False
制造业增加值                 False
建筑业增加值                 False
批发和零售业增加值               False
交通运输、仓储和邮政业增加值          False
住宿和餐饮业增加值               False
金融业增加值                 False
房地产业增加值                False
信息传输、软件和信息技术服务业增加值     False
租赁和商务服务业增加值              False
其他行业增加值                False
dtype: bool
countryName               False
province_confirmedCount   False
province_curedCount       False
province_deadCount        False
dtype: bool
title      False
summary    False
dtype: bool
```

In [8]:

```
import matplotlib.pyplot as plt
import matplotlib
import os

%matplotlib inline
# 指定中文字体
fpath = os.path.join(r"C:\Users\崔梦茹\Documents\作业\数据技术分析技术作业/NotoSansCJK.otf")
myfont = matplotlib.font_manager.FontProperties(fname=fpath)
# 绘图
data_world = data_world.sort_values(by='确诊人数', ascending=False)  # 按确诊人数进行排序
data_world_set = data_world[['确诊人数', '治愈人数', '死亡人数']]
data_world_set.index = data_world['国家名称']
data_world_set.head(10).plot(kind='bar', figsize=(15, 10))  # 对排序前十的国家数据进行绘图
plt.xlabel('国家名称', fontproperties=myfont)
plt.xticks(fontproperties=myfont)
plt.legend(fontsize=30, prop=myfont)  # 设置图例
```

Out[8]:

<matplotlib.legend.Legend at 0x1c55be085b0>



In [9]:

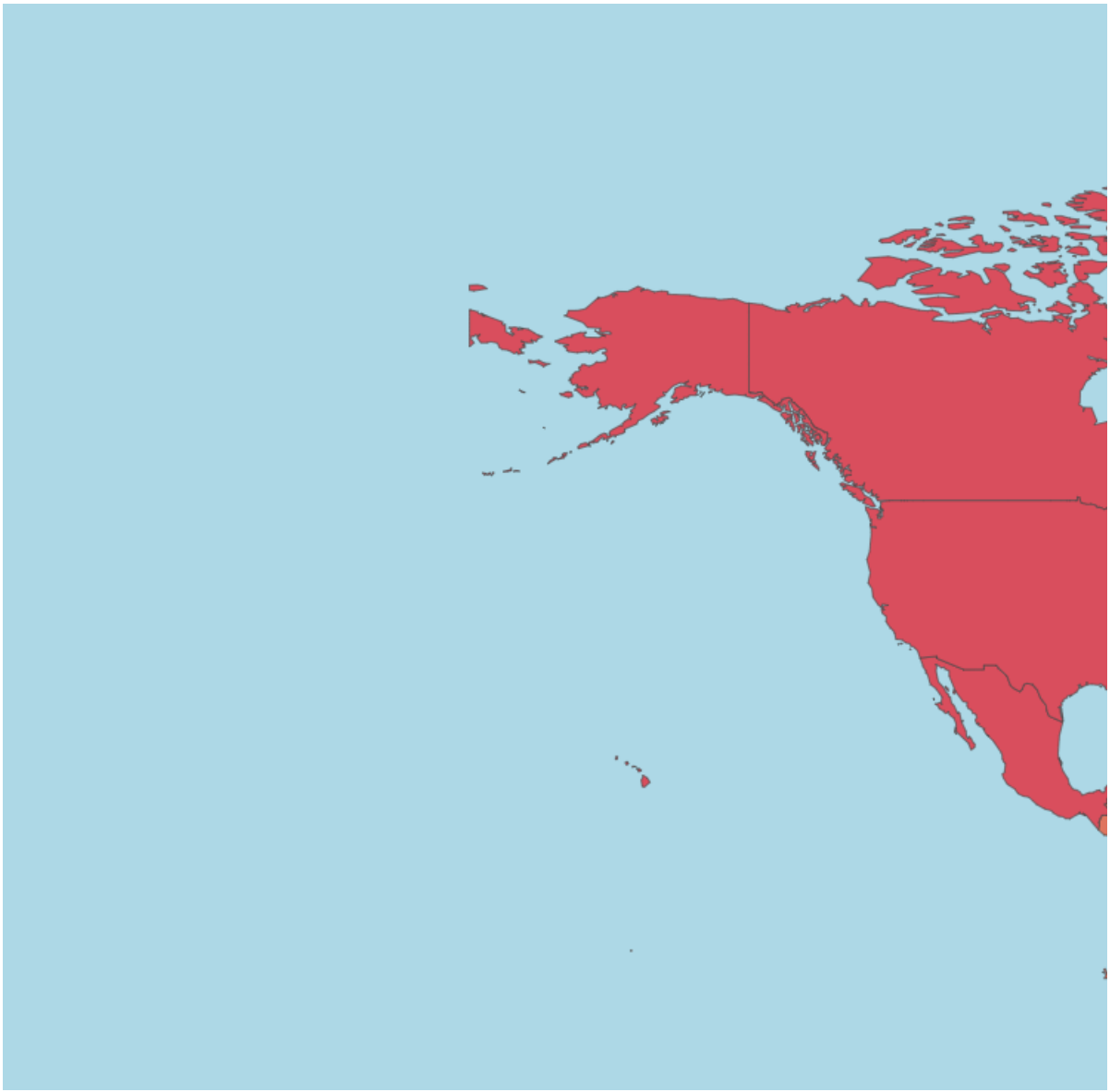
```
from pyecharts.charts import Map
from pyecharts import options as opts
from pyecharts.globals import CurrentConfig, NotebookType

CurrentConfig.NOTEBOOK_TYPE = NotebookType.JUPYTER_NOTEBOOK
name_map = {  # 世界各国数据的中英文对比
    'Singapore Rep.': '新加坡',
    'Dominican Rep.': '多米尼加',
    'Palestine': '巴勒斯坦',
    'Bahamas': '巴哈马',
    'Timor-Leste': '东帝汶',
    'Afghanistan': '阿富汗',
    'Guinea-Bissau': '几内亚比绍',
    "Côte d'Ivoire": '科特迪瓦',
    'Siachen Glacier': '锡亚琴冰川',
    "Br. Indian Ocean Ter.": '英属印度洋领土',
    'Angola': '安哥拉',
    'Albania': '阿尔巴尼亚',
    'United Arab Emirates': '阿联酋',
    'Argentina': '阿根廷',
    'Armenia': '亚美尼亚',
    'French Southern and Antarctic Lands': '法属南半球和南极领地',
    'Australia': '澳大利亚',
    'Austria': '奥地利',
    'Azerbaijan': '阿塞拜疆',
    'Burundi': '布隆迪',
    'Belgium': '比利时',
    'Benin': '贝宁',
    'Burkina Faso': '布基纳法索',
    'Bangladesh': '孟加拉国',
    'Bulgaria': '保加利亚',
    'The Bahamas': '巴哈马',
    'Bosnia and Herz.': '波斯尼亚和黑塞哥维那',
    'Belarus': '白俄罗斯',
    'Belize': '伯利兹',
    'Bermuda': '百慕大',
    'Bolivia': '玻利维亚',
    'Brazil': '巴西',
    'Brunei': '文莱',
    'Bhutan': '不丹',
    'Botswana': '博茨瓦纳',
    'Central African Rep.': '中非',
    'Canada': '加拿大',
    'Switzerland': '瑞士',
    'Chile': '智利',
    'China': '中国',
    'Ivory Coast': '象牙海岸',
    'Cameroon': '喀麦隆',
    'Dem. Rep. Congo': '刚果民主共和国',
    'Congo': '刚果',
    'Colombia': '哥伦比亚',
    'Costa Rica': '哥斯达黎加',
    'Cuba': '古巴',
    'N. Cyprus': '北塞浦路斯',
    'Cyprus': '塞浦路斯',
    'Czech Rep.': '捷克',
    'Germany': '德国',
    'Djibouti': '吉布提',
    'Denmark': '丹麦',
```

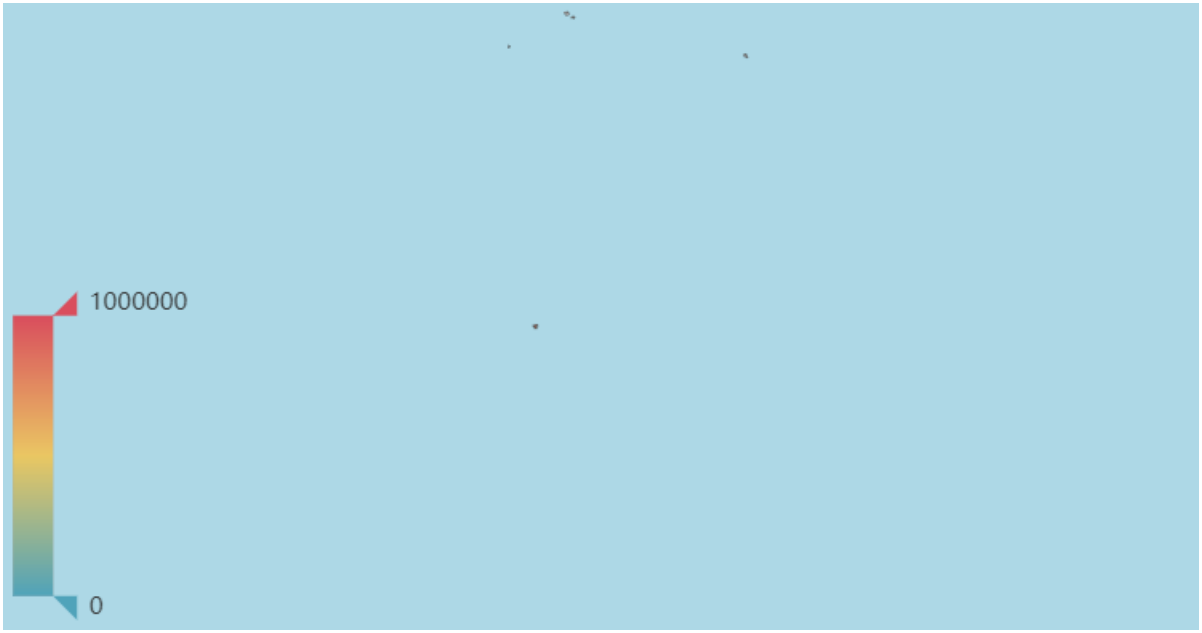
```
    'Algeria': '阿尔及利亚',
    'Ecuador': '厄瓜多尔',
    'Egypt': '埃及',
    'Eritrea': '厄立特里亚',
    'Spain': '西班牙',
    'Estonia': '爱沙尼亚',
    'Ethiopia': '埃塞俄比亚',
    'Finland': '芬兰',
    'Fiji': '斐',
    'Falkland Islands': '福克兰群岛',
    'France': '法国',
    'Gabon': '加蓬',
    'United Kingdom': '英国',
    'Georgia': '格鲁吉亚',
    'Ghana': '加纳',
    'Guinea': '几内亚',
    'Gambia': '冈比亚',
    'Guinea Bissau': '几内亚比绍',
    'Eq. Guinea': '赤道几内亚',
    'Greece': '希腊',
    'Greenland': '格陵兰',
    'Guatemala': '危地马拉',
    'French Guiana': '法属圭亚那',
    'Guyana': '圭亚那',
    'Honduras': '洪都拉斯',
    'Croatia': '克罗地亚',
    'Haiti': '海地',
    'Hungary': '匈牙利',
    'Indonesia': '印度尼西亚',
    'India': '印度',
    'Ireland': '爱尔兰',
    'Iran': '伊朗',
    'Iraq': '伊拉克',
    'Iceland': '冰岛',
    'Israel': '以色列',
    'Italy': '意大利',
    'Jamaica': '牙买加',
    'Jordan': '约旦',
    'Japan': '日本',
    'Kazakhstan': '哈萨克斯坦',
    'Kenya': '肯尼亚',
    'Kyrgyzstan': '吉尔吉斯斯坦',
    'Cambodia': '柬埔寨',
    'Korea': '韩国',
    'Kosovo': '科索沃',
    'Kuwait': '科威特',
    'Lao PDR': '老挝',
    'Lebanon': '黎巴嫩',
    'Liberia': '利比里亚',
    'Libya': '利比亚',
    'Sri Lanka': '斯里兰卡',
    'Lesotho': '莱索托',
    'Lithuania': '立陶宛',
    'Luxembourg': '卢森堡',
    'Latvia': '拉脱维亚',
    'Morocco': '摩洛哥',
    'Moldova': '摩尔多瓦',
    'Madagascar': '马达加斯加',
    'Mexico': '墨西哥',
    'Macedonia': '马其顿',
    'Mali': '马里',
```

```
    'Myanmar': '缅甸',
    'Montenegro': '黑山',
    'Mongolia': '蒙古',
    'Mozambique': '莫桑比克',
    'Mauritania': '毛里塔尼亚',
    'Malawi': '马拉维',
    'Malaysia': '马来西亚',
    'Namibia': '纳米比亚',
    'New Caledonia': '新喀里多尼亚',
    'Niger': '尼日尔',
    'Nigeria': '尼日利亚',
    'Nicaragua': '尼加拉瓜',
    'Netherlands': '荷兰',
    'Norway': '挪威',
    'Nepal': '尼泊尔',
    'New Zealand': '新西兰',
    'Oman': '阿曼',
    'Pakistan': '巴基斯坦',
    'Panama': '巴拿马',
    'Peru': '秘鲁',
    'Philippines': '菲律宾',
    'Papua New Guinea': '巴布亚新几内亚',
    'Poland': '波兰',
    'Puerto Rico': '波多黎各',
    'Dem. Rep. Korea': '朝鲜',
    'Portugal': '葡萄牙',
    'Paraguay': '巴拉圭',
    'Qatar': '卡塔尔',
    'Romania': '罗马尼亚',
    'Russia': '俄罗斯',
    'Rwanda': '卢旺达',
    'W. Sahara': '西撒哈拉',
    'Saudi Arabia': '沙特阿拉伯',
    'Sudan': '苏丹',
    'S. Sudan': '南苏丹',
    'Senegal': '塞内加尔',
    'Solomon Is.': '所罗门群岛',
    'Sierra Leone': '塞拉利昂',
    'El Salvador': '萨尔瓦多',
    'Somaliland': '索马里兰',
    'Somalia': '索马里',
    'Serbia': '塞尔维亚',
    'Suriname': '苏里南',
    'Slovakia': '斯洛伐克',
    'Slovenia': '斯洛文尼亚',
    'Sweden': '瑞典',
    'Swaziland': '斯威士兰',
    'Syria': '叙利亚',
    'Chad': '乍得',
    'Togo': '多哥',
    'Thailand': '泰国',
    'Tajikistan': '塔吉克斯坦',
    'Turkmenistan': '土库曼斯坦',
    'East Timor': '东帝汶',
    'Trinidad and Tobago': '特里尼达和多巴哥',
    'Tunisia': '突尼斯',
    'Turkey': '土耳其',
    'Tanzania': '坦桑尼亚',
    'Uganda': '乌干达',
    'Ukraine': '乌克兰',
    'Uruguay': '乌拉圭',
```

```
    'United States': '美国',
    'Uzbekistan': '乌兹别克斯坦',
    'Venezuela': '委内瑞拉',
    'Vietnam': '越南',
    'Vanuatu': '瓦努阿图',
    'West Bank': '西岸',
    'Yemen': '也门',
    'South Africa': '南非',
    'Zambia': '赞比亚',
    'Zimbabwe': '津巴布韦',
    'Comoros': '科摩罗'
}

map = Map(init_opts=opts.InitOpts(width="1900px", height="900px",
                                  bg_color="#ADD8E6", page_title="全球疫情确诊人数"))  # 获得世界地
map.add("确诊人数", [list(z) for z in zip(data_world['国家名称'], data_world['确诊人数'])],
        is_map_symbol_show=False,  # 添加确诊人数信息
        # 通过name_map来转化国家的中英文名称方便显示
        maptype="world", label_opts=opts.LabelOpts(is_show=False), name_map=name_map,
        itemstyle_opts=opts.ItemStyleOpts(color="rgb(49,60,72)"),
        ).set_global_opts(
    visualmap_opts=opts.VisualMapOpts(max_=1000000),  # 对视觉映射进行配置
)
map.render_notebook()  # 在notebook中显示
```

Out[9]:





In [10]:

```
country = data_area_times.sort_values('province_confirmedCount', ascending=False).drop_duplicates(
    subset='countryName', keep='first').head(6)['countryName']
country = list(country)  # 对于同一天采集的多个数据，只保留第一次出现的数据也就是最后一次更新的数据
country
```

Out[10]:

['美国', '巴西', '印度', '俄罗斯', '秘鲁', '智利']

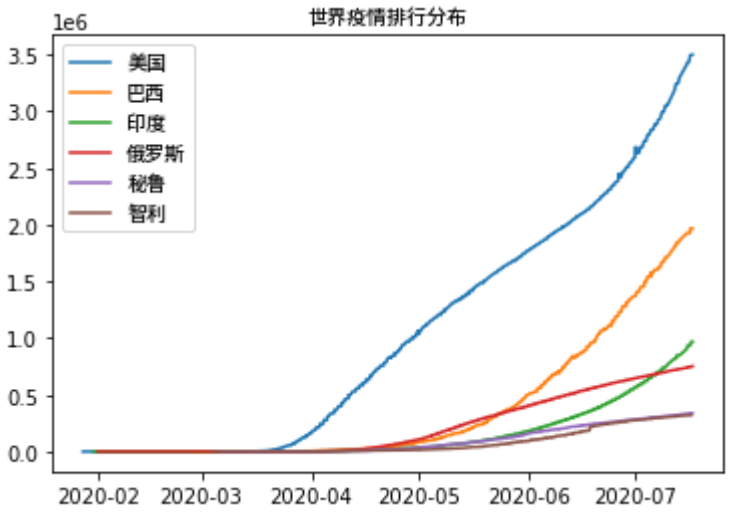
In [11]:

```
data_America = data_area_times[data_area_times['countryName'] == '美国']
data_Brazil = data_area_times[data_area_times['countryName'] == '巴西']
data_India = data_area_times[data_area_times['countryName'] == '印度']
data_Russia = data_area_times[data_area_times['countryName'] == '俄罗斯']
data_Peru = data_area_times[data_area_times['countryName'] == '秘鲁']
data_Chile = data_area_times[data_area_times['countryName'] == '智利']

timeindex = data_area_times.index
timeindex = timeindex.floor('D')  # 对于日期索引，只保留具体到哪一天
data_area_times.index = timeindex

timeseries = pd.DataFrame(data_America.index)
timeseries.index = data_America.index
data_America = pd.concat([timeseries, data_America], axis=1)
data_America.drop_duplicates(
    subset='updateTime', keep='first', inplace=True)  # 对美国数据进行处理，获得美国确诊人数的时间序
data_America.drop('updateTime', axis=1, inplace=True)

timeseries = pd.DataFrame(data_Brazil.index)
timeseries.index = data_Brazil.index
data_Brazil = pd.concat([timeseries, data_Brazil], axis=1)
# 对巴西数据进行处理，获得巴西确诊人数的时间序列
data_Brazil.drop_duplicates(subset='updateTime', keep='first', inplace=True)
data_Brazil.drop('updateTime', axis=1, inplace=True)

timeseries = pd.DataFrame(data_India.index)
timeseries.index = data_India.index
data_India = pd.concat([timeseries, data_India], axis=1)
# 对印度数据进行处理，获得印度确诊人数的时间序列
data_India.drop_duplicates(subset='updateTime', keep='first', inplace=True)
data_India.drop('updateTime', axis=1, inplace=True)

timeseries = pd.DataFrame(data_Russia.index)
timeseries.index = data_Russia.index
data_Russia = pd.concat([timeseries, data_Russia], axis=1)
# 对俄罗斯数据进行处理，获得俄罗斯确诊人数的时间序列
data_Russia.drop_duplicates(subset='updateTime', keep='first', inplace=True)
data_Russia.drop('updateTime', axis=1, inplace=True)

timeseries = pd.DataFrame(data_Peru.index)
timeseries.index = data_Peru.index
data_Peru = pd.concat([timeseries, data_Peru], axis=1)
# 对秘鲁数据进行处理，获得秘鲁确诊人数的时间序列
data_Peru.drop_duplicates(subset='updateTime', keep='first', inplace=True)
data_Peru.drop('updateTime', axis=1, inplace=True)

timeseries = pd.DataFrame(data_Chile.index)
timeseries.index = data_Chile.index
data_Chile = pd.concat([timeseries, data_Chile], axis=1)
# 对智利数据进行处理，获得智利确诊人数的时间序列
data_Chile.drop_duplicates(subset='updateTime', keep='first', inplace=True)
data_Chile.drop('updateTime', axis=1, inplace=True)

plt.title("世界疫情排行分布", fontproperties=myfont)
plt.plot(data_America['province_confirmedCount'])
plt.plot(data_Brazil['province_confirmedCount'])
plt.plot(data_India['province_confirmedCount'])
plt.plot(data_Russia['province_confirmedCount'])
plt.plot(data_Peru['province_confirmedCount'])
```

```
plt.plot(data_Chile['province_confirmedCount'])
plt.legend(country, prop=myfont)
```

Out[11]:

```
<matplotlib.legend.Legend at 0x1c56300bca0>
```



In [12]:

```
pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\anaconda\lib\site-packages (1.8.1)
Requirement already satisfied: numpy>=1.6.1 in c:\anaconda\lib\site-packages (from w
ordcloud) (1.20.3)
Requirement already satisfied: pillow in c:\anaconda\lib\site-packages (from wordclo
ud) (8.4.0)
Requirement already satisfied: matplotlib in c:\anaconda\lib\site-packages (from wor
dcloud) (3.4.3)
Requirement already satisfied: cycler>=0.10 in c:\anaconda\lib\site-packages (from m
atplotlib->wordcloud) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\anaconda\lib\site-packages (f
rom matplotlib->wordcloud) (1.3.1)
Requirement already satisfied: python-dateutil>=2.7 in c:\anaconda\lib\site-packages
(from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in c:\anaconda\lib\site-packages (fr
om matplotlib->wordcloud) (3.0.4)
Requirement already satisfied: six in c:\anaconda\lib\site-packages (from cycler>=0.
10->matplotlib->wordcloud) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [13]:

```
pip install jieba
```

```
Requirement already satisfied: jieba in c:\anaconda\lib\site-packages (0.42.1)Note:
you may need to restart the kernel to use updated packages.
```

In [14]:

```
import jieba
import re
from wordcloud import WordCloud

def word_cut(x): return jieba.lcut(x)  # 进行结巴分词

news = []
reg = "[^\u4e00-\u9fa5]"
for i in data_news['title']:
    if re.sub(reg, '', i) != '':  # 去掉英文数字和标点等无关字符，仅保留中文词组
        news.append(re.sub(reg, '', i))  # 用news列表汇总处理后的新闻标题

words = []
counts = {}
for i in news:
    words.append(word_cut(i))  # 对所有新闻进行分词
for word in words:
    for a_word in word:
        if len(a_word) == 1:
            continue
        else:
            counts[a_word] = counts.get(a_word, 0)+1  # 用字典存储对应分词的词频
words_sort = list(counts.items())
words_sort.sort(key=lambda x: x[1], reverse=True)

newcloud = WordCloud(font_path=r"C:\Users\崔梦茹\Documents\作业\数据技术分析技术作业\NotoSansCJK.otf
                     background_color="white", width=600, height=300, max_words=50)  # 生成词云
newcloud.generate_from_frequencies(counts)
image = newcloud.to_image()  # 转换成图片
image
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\崔梦茹\AppData\Local\Temp\jieba.cache
Loading model cost 0.981 seconds.
Prefix dict has been built successfully.
```

Out[14]:

In [15]:

```
pip install -i https://pypi.douban.com/simple gensim
```

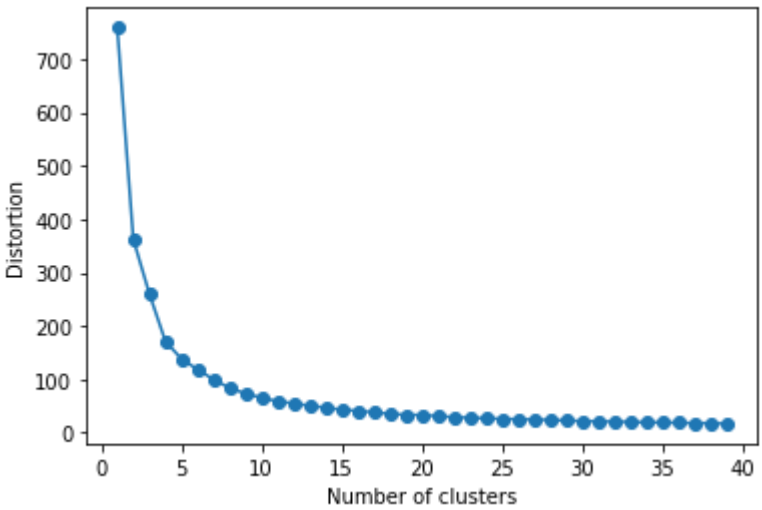
```
Looking in indexes: https://pypi.douban.com/simple (https://pypi.douban.com/simple)
Requirement already satisfied: gensim in c:\anaconda\lib\site-packages (4.2.0)
Requirement already satisfied: smart-open>=1.8.1 in c:\anaconda\lib\site-packages (f
rom gensim) (6.0.0)
Requirement already satisfied: Cython==0.29.28 in c:\anaconda\lib\site-packages (fro
m gensim) (0.29.28)
Requirement already satisfied: numpy>=1.17.0 in c:\anaconda\lib\site-packages (from
 gensim) (1.20.3)
Requirement already satisfied: scipy>=0.18.1 in c:\anaconda\lib\site-packages (from
 gensim) (1.7.1)
Note: you may need to restart the kernel to use updated packages.
```

In [16]:

```
from gensim.models import Word2Vec
from sklearn.cluster import KMeans
import warnings
warnings.filterwarnings('ignore')

words = []

for i in news:
    words.append(word_cut(i))
model = Word2Vec(words, sg=0, vector_size=300, window=5, min_count=5)
keys = model.wv.key_to_index.keys()
wordvector = []
for key in keys:
    wordvector.append(model.wv[key])

distortions = []
for i in range(1, 40):
    word_kmeans = KMeans(n_clusters=i,
                         init='k-means++',
                         n_init=10,
                         max_iter=300,
                         random_state=0)
    word_kmeans.fit(wordvector)
    distortions.append(word_kmeans.inertia_)
plt.plot(range(1, 40), distortions, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Distortion')
```

Out[16]:

```
Text(0, 0.5, 'Distortion')
```



In [17]:

```
word_kmeans = KMeans(n_clusters=10)  # 聚成10类
word_kmeans.fit(wordvector)

labels = word_kmeans.labels_

for num in range(0, 10):
    text = []
    for i in range(len(keys)):
        if labels[i] == num:
            text.append(list(keys)[i])  # 分别获得10类的聚类结果
    print(text)
```

```
['湖南', '贵州', '河北', '山西', '青海', '重庆市', '天津市', '黑龙江省', '墨西
哥', '吉林省', '比利时', '印尼', '沙特', '云南省', '澳门', '摩洛哥', '省区市',
'瑞士', '贵州省', '安徽省', '以外', '河南省', '发布会', '卫健委日', '超例', '辽宁
省', '刚果', '哈萨克斯坦', '英国首相', '内蒙古自治区', '匈牙利', '全区', '黎巴
嫩', '俄', '绥芬河', '集体', '实验室', '家', '金银', '首个', '趋缓', '研发', '员
工', '多名', '业务', '经验', '病人', '吉林市', '受新冠', '升级', '救治', '今天',
'月底', '调查', '快速', '放松', '封闭', '但', '小汤山', '外', '门诊', '世界卫生组
织', '状态', '主席', '需要', '治愈率', '火神', '两例', '失业', '破', '山', '降',
'乘客', '出行', '行动', '预约', '群体', '轨迹', '做', '重要', '重点', '逼近',
'零', '逾', '联合国', '启程', '军队', '有序', '一名', '失业率', '监狱', '采样',
'生活', '万份', '生产', '第二阶段', '诊疗', '公共', '市民', '复苏', '部长', '用',
'省', '撤离', '停止', '总', '一季度', '临床', '恶化', '民航局', '规定', '大部分',
'至人', '下降', '包机', '总干事', '病毒感染', '阶段', '采取', '购买', '药物', '返
京', '密接', '中国政府', '欧盟', '试剂', '今起', '前往', '都', '流行', '提高',
'方式', '航线', '现', '奥组委', '交易', '万人次', '安排', '尚未', '力度', '市',
'厄瓜多尔', '一周', '收治', '个人', '食品', '市长', '采购', '最小', '床位', '抗
议', '暂', '哈尔滨', '央视', '期', '外籍', '推动', '今年', '无法', '旅游业', '野
生动物', '补助', '落实', '两天', '名新冠', '神山', '隐瞒', '警告', '裁员', '再
度', '连降', '世界', '任务', '排除', '院士', '诊断', '项目', '外长', '缓解', '返
回', '举措', '高', '收到', '主要', '传染病', '运营', '因新冠', '针对', '海鲜',
'日前', '处于', '加大', '副', '不是', '吨', '区域', '召开', '迪拜', '大区', '万
个', '基本', '合肥', '撤侨', '当地', '中心', '增幅', '试剂盒', '援', '鲍里斯',
'禁令', '每天', '亿只', '临床试验', '呼吸机', '提升', '秘书', '餐厅', '举办', '封
城', '人民', '价格', '分享', '使馆', '自', '同时', '鄂', '城市', '二级', '行业',
'高三', '以下', '等国', '扩散', '引', '两', '流感', '监护', '团队', '援鄂', '创
新', '社交', '可以', '加剧', '坚决', '省份', '万名', '航空', '半数', '事态', '告
急', '其他', '总领馆', '资金', '降至例', '使用', '严防', '证据', '爱心', '系',
'籍', '不明', '急', '运输', '万次', '佛罗里达州', '吗', '认为', '三级', '保持',
'启用', '此前', '乘', '复学', '移动', '过去', '代表', '免疫', '即将', '入院', '建
设', '说', '恢复正常', '首相', '防护', '多项', '捐款', '氯喹', '羟', '财政', '任
何', '准备', '万多', '保护', '执行', '领导', '首尔', '大臣', '全力', '现在', '机
制', '若', '正常', '等级', '夜店', '沪', '人士', '网络', '也', '首日', '宣言',
'急需', '迎接', '让', '工资', '英雄', '逝者', '开', '解禁', '禁足', '追加', '毕业
生', '接待', '经', '转机', '体温', '欧元', '联合', '变化', '幼儿园', '酒吧', '复
阳', '紧张', '州长', '封闭式', '投资', '波', '包括', '主流', '环境', '公务员',
'出租车', '发']
['最新', '巴西', '年月日', '日时', '泰国', '新加坡', '卫健委', '性', '截至',
'均', '北京市', '专家组', '从', '达', '清零', '于', '医护人员', '最大', '好消息',
'持续', '年', '非洲', '取消', '实施', '全部', '目前', '前', '今日', '受', '者',
'首都', '国', '计划', '总理', '东京', '万人', '特朗普', '要求', '一级', '公主',
'居家', '近', '活动', '小时', '上升', '举行', '最', '提供', '岁', '关闭', '卫生',
'重启', '新闻', '企业', '因', '委员会', '完成', '合作', '观察', '阳性', '开始',
'公民', '援助', '全面', '发地', '内', '佩戴', '暴发', '解除', '相关', '支持', '最
高', '美', '捐赠', '及', '封锁', '牺牲', '抵达', '市场', '继续', '钟南山', '首
批', '研究', '时间', '再次', '发生', '应', '同胞', '来', '应急', '大', '医生',
'戴', '解封', '支援', '调整', '要', '逝世', '五一', '儿童', '管理', '发热', '呼
```

吁', '服务', '复课', '呈', '逐步', '成', '推迟', '数据', '海外', '延期', '到', '日起', '社会', '没有', '假期', '开展', '面临', '大使馆', '奥运会', '社区', '加强', '扩大', '显示', '实行', '约', '并', '回国', '亿', '烈士', '提醒', '出台', '以上', '高校', '约翰逊', '事件', '学校', '队员', '景区']

['月', '冠状病毒', '美国', '北京', '情况', '感染', '患者', '湖北', '人', '全国', '为', '出现', '至', '人数', '名', '起', '连续', '宣布', '对', '万', '或', '个', '被', '发布', '天', '公布', '首次', '工作', '所有', '影响', '新', '又', '开放', '专家', '启动', '一', '经济', '有', '发现', '聚集', '进行', '应对', '是', '等', '国际', '多', '风险', '地区', '后', '中', '医疗队', '响应', '驻', '再', '称', '了', '总统', '可能', '限制', '进入', '复工', '部分', '可', '增加', '向', '丨', '号', '欧洲', '暂停', '健康', '病毒检测', '民众', '令', '开学', '重症', '疫苗', '传播', '政府', '航班', '紧急', '国内', '仍', '抗体', '返校', '入境', '悼念', '旅客']

['例', '确诊', '新增', '病例', '肺炎', '累计']

['荷兰', '阿根廷', '福建省', '四川省', '巴基斯坦', '葡萄牙', '江西省', '山西省', '波兰', '智利', '新西兰', '希腊', '肯尼亚', '青海省', '哥伦比亚', '河北省', '斯洛伐克', '塞尔维亚', '巴林', '乌克兰', '共计', '浙江省', '以色列', '伊拉克', '孟加拉国', '湖南省', '贫民窟', '塞内加尔', '白俄罗斯', '详情', '赞比亚', '捷克', '例均', '破万', '西藏自治区', '格鲁吉亚', '科威特', '毛里求斯', '境内', '越南', '奥地利', '缅甸', '斯里兰卡', '乌兹别克斯坦', '卢森堡', '总数', '至时', '江苏省', '纽约市', '卫健委月', '阿塞拜疆', '台湾', '突尼斯', '压力', '金', '考虑', '证明', '情况通报', '阿曼', '婴儿', '辽宁大连', '上调', '白宫', '利比亚', '雷', '罚款', '保加利亚', '文莱', '卡塔尔', '喀麦隆', '默哀', '阿尔巴尼亚', '派', '很', '教育部', '沈阳', '千例', '岗位', '布', '牡丹江', '危重', '给', '所', '考试', '统计', '得到', '会议', '生命', '埃塞俄比亚', '超人', '化', '亚美尼亚', '通告', '有关', '倍', '好', '严峻', '住院', '一个月', '工作人员', '下跌', '展开', '春节假期', '筛查', '发言人', '迪士尼', '复航', '流动', '返程', '家中', '海滩', '多地', '量', '也门', '吴尊友', '上海市', '指导', '舒兰市', '停运', '停课', '进出', '作用', '水平', '检疫', '津巴布韦', '型', '全员', '三个', '数量', '死于', '记者', '滞留', '立陶宛', '安道尔', '行程', '加纳', '洛杉矶', '多家', '减少', '之下', '男子', '重开', '黄石', '危机', '北美', '一例', '接近', '深圳', '普京', '例例', '第二批', '预测', '乌拉圭', '吉布提', '史', '圭亚那', '病情', '尼日利亚', '轻症', '错峰', '近万人', '叙利亚', '各', '销售', '胜利', '警惕', '发改委', '多国', '柳叶刀', '武汉协和医院', '阿尔及利亚', '营业', '回', '严禁', '供应链', '实现', '上班', '南京', '出征', '凯旋', '订正', '布基纳法索', '资助', '第一批', '蒙古国', '至例', '增例', '坦桑尼亚', '老挝', '塞浦路斯', '有名', '襄阳', '构成', '堂食', '同一', '陆续', '日本政府', '加快', '明确', '补贴', '办理', '而', '过万', '医务', '引发', '苏丹', '网友', '仪式', '两万', '不足', '国际航班', '传染', '亚洲', '接收', '比赛', '吉尔吉斯斯坦', '挑战', '数超', '序列', '名单', '近万', '外卖', '比', '座', '正在', '案例', '多数', '纳入', '明显', '圈', '经济衰退', '规模', '疾控', '削减', '具备', '级', '主任', '出席', '发出', '冠', '有效', '两个', '酒店', '地方', '关键', '心理', '大厅', '疾病', '须', '帮助', '指南', '这些', '老人', '如何', '办事处', '视频', '省市', '民航', '不断', '工人', '系统', '参与', '强调', '分批', '赤道几内亚', '供应', '近例', '大会', '次', '啦', '航空公司', '马里', '就诊', '份', '尚', '条', '建', '蛋白质', '全', '回升', '团结', '全体', '参加', '中考', '变', '看', '其', '防止', '处以', '多州', '进京']

['疫情', '的', '将', '中国', '武汉', '病毒', '检测', '国家', '和', '在', '防控', '抗疫', '医院', '口罩', '措施', '隔离', '核酸', '已', '人员', '医疗', '恢复', '不', '物资', '延长', '与', '防疫', '期间']

['万例', '全球', '世卫', '组织', '超过', '超']

['新冠', '例新冠', '日', '出院', '无', '输入', '境外', '治愈', '通报', '报告', '本地']

['新型', '死亡', '上海', '首例', '天津', '日本', '增至', '英国', '意大利', '韩国', '达例', '感染者', '西班牙', '升至', '印度', '广东', '德国', '辽宁', '时', '黑龙江', '山东', '重庆', '广西', '无症状', '香港', '俄罗斯', '四川', '陕西', '法国', '超万', '伊朗', '江苏', '福建', '云南', '疑似病例', '安徽', '浙江', '甘肃', '单日', '内蒙古', '吉林', '新疆', '海南', '加拿大', '天无', '昨日', '有例', '本土', '其中']

['河南', '江西', '宁夏', '澳大利亚', '菲律宾', '土耳其', '山东省', '马来西亚', '广东省', '湖北省', '突破', '直播', '阴性', '第例', '来自', '紧急状态', '西藏', '秘鲁', '年月日时', '例为', '阿联酋', '机构', '纽约州', '医学观察', '严重', '达

到', '管控', '钻石', '强制', '疑似', '抗击', '康复', '确认', '含', '南非', '外交部', '重新', '兵团', '共有', '现有', '埃及', '结束', '以来', '已经', '赴', '集中', '会', '已有', '接受', '高风险', '一个', '治疗', '能力', '驰援', '进一步', '复产', '全省', '卫生部长', '至少', '反弹', '方舱', '邮轮', '数', '安全', '医务人员', '建议', '一天', '疾控中心', '结果', '共', '增长', '控制', '信息', '机场', '允许', '去世', '亿元', '卫生部', '广州', '禁止', '曾', '加州', '密切接触', '回应', '蔓延', '放宽', '医用', '高考', '需', '地', '养老院', '推出', '留学生', '关联', '痊愈', '护士', '武汉市', '口岸', '免费', '正式', '医护', '官员', '旅游', '学生', '正', '国务院', '发放', '工作者', '预计', '小区', '问题', '各地', '各国', '感谢', '冲击', '范围', '人群', '两周', '场所', '最后', '上', '一线', '潭', '加速', '下周', '临时', '运抵', '旅行', '边境', '关于', '我', '通过', '居民', '一律', '低', '每日', '以', '积极', '公共卫生', '铁路', '中小学', '除', '莫斯科', '张文宏', '症状', '媒体', '公司', '非', '客运', '消费', '我国', '不会', '排查', '导致', '接触', '纽约', '蔬菜', '占', '州', '未来', '中央', '常态', '第二', '决定', '必须', '突发', '发展', '成为', '黄冈', '下', '大规模', '高峰', '是否', '月份', '中方', '注意', '家庭', '深切', '回家', '下半旗', '宵禁', '表示', '游客', '通知', '未', '还', '级别', '医疗机构', '助力', '批准', '战疫', '确定', '样本', '就', '救助', '外出', '快递', '就业', '不得', '由', '年级', '政策', '形势', '大幅', '部门', '做好', '测试', '关注', '更', '共同', '师生', '申请', '保障', '同比', '严格', '只', '志哀', '线上', '联防', '联控', '一年', '避免', '官方', '考生', '下调', '距离', '表明']

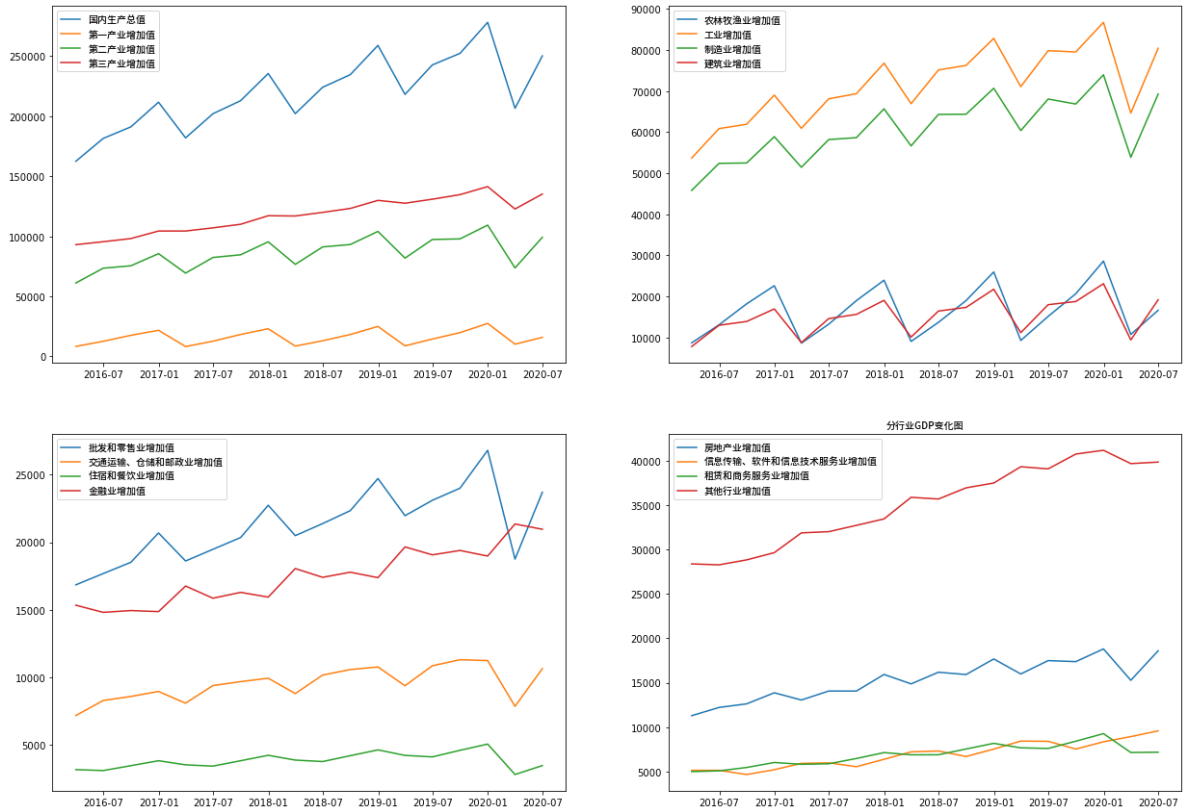
In [18]:

```
sum_GDP = ['国内生产总值', '第一产业增加值', '第二产业增加值', '第三产业增加值']
industry_GDP = ['农林牧渔业增加值', '工业增加值', '制造业增加值', '建筑业增加值']
industry2_GDP = ['批发和零售业增加值', '交通运输、仓储和邮政业增加值', '住宿和餐饮业增加值', '金融业
industry3_GDP = ['房地产业增加值', '信息传输、软件和信息技术服务业增加值',
                 '租赁和商务服务业增加值', '其他行业增加值']  # 对不同行业分四类来展现

fig = plt.figure()
fig, axes = plt.subplots(2, 2, figsize=(21, 15))  # 分别用四个子图来展现数据变化情况

axes[0][0].plot(data_economy[sum_GDP])
axes[0][0].legend(sum_GDP, prop=myfont)
axes[0][1].plot(data_economy[industry_GDP])
axes[0][1].legend(industry_GDP, prop=myfont)
axes[1][0].plot(data_economy[industry2_GDP])
axes[1][0].legend(industry2_GDP, prop=myfont)
axes[1][1].plot(data_economy[industry3_GDP])
axes[1][1].legend(industry3_GDP, prop=myfont)

plt.title('分行业GDP变化图', fontproperties=myfont)
```

Out[18]:

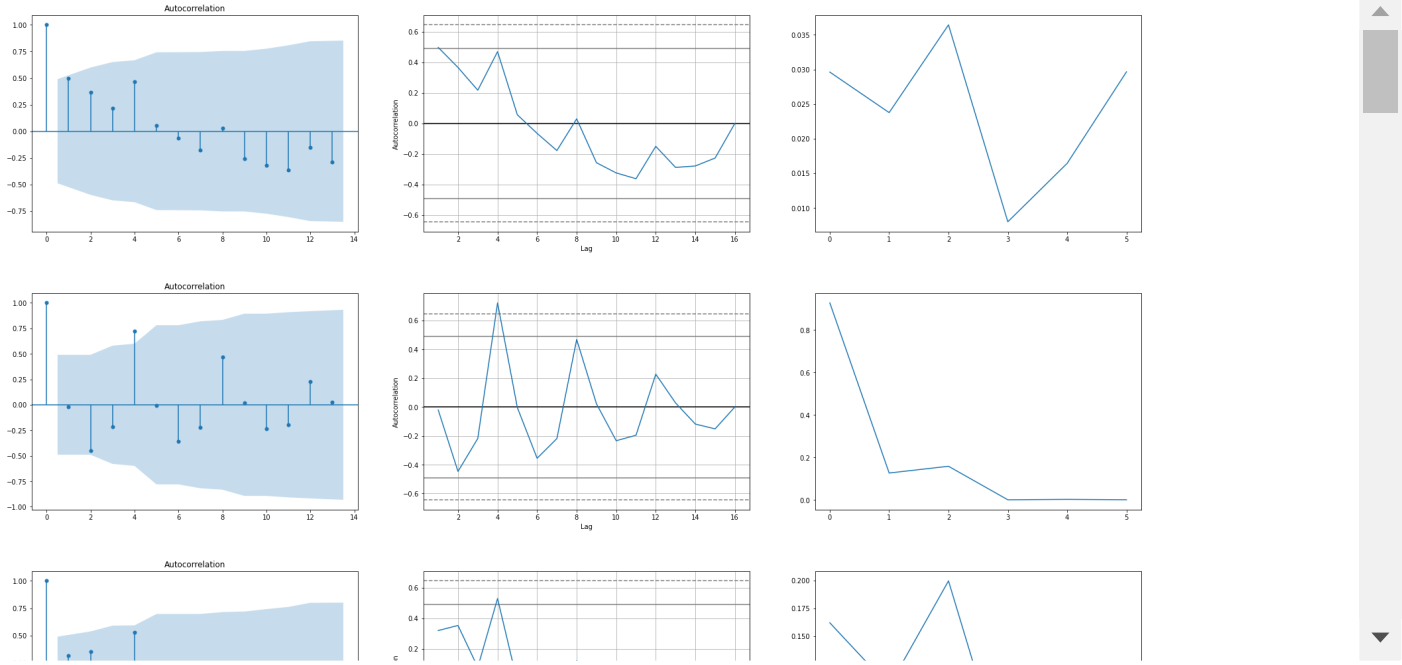
Text(0.5, 1.0, '分行业GDP变化图')

<Figure size 432x288 with 0 Axes>



In [19]:

```
from statsmodels.graphics.tsaplots import plot_acf
from pandas.plotting import autocorrelation_plot
from statsmodels.sandbox.stats.diagnostic import acorr_ljungbox


GDP_type = ['国内生产总值', '第一产业增加值', '第二产业增加值', '第三产业增加值',
            '农林牧渔业增加值', '工业增加值', '制造业增加值', '建筑业增加值', '批发和零售业增加值',
            '交通运输、仓储和邮政业增加值', '住宿和餐饮业增加值', '金融业增加值',
            '房地产业增加值', '信息传输、软件和信息技术服务业增加值', '租赁和商务服务业增加值', '其他

for i in GDP_type:
    each_data = data_economy[i][:-2]
    plt.figure(figsize=(30, 6))
    ax1 = plt.subplot(1, 3, 1)
    ax2 = plt.subplot(1, 3, 2)
    ax3 = plt.subplot(1, 3, 3)
    LB2, P2 = acorr_ljungbox(each_data)  # 进行纯随机性检验
    plot_acf(each_data, ax=ax1)
    autocorrelation_plot(each_data, ax=ax2)  # 进行平稳性检验
    ax3.plot(P2)
```



In [20]:

```
from statsmodels.tsa.arima_model import ARMA
from statsmodels.tsa.stattools import arma_order_select_ic

warnings.filterwarnings('ignore')
data_arma = pd.DataFrame(data_economy['国内生产总值'][:-2])  # 选取疫情期前的16个季度进行建模
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()  # 使用ARMA建模
rate1 = list(data_economy['国内生产总值'][-2] /
             arma.forecast(steps=1)[0])  # 获得疫情期当季度的预测值
rate1  # 实际值与预测值的比率
```

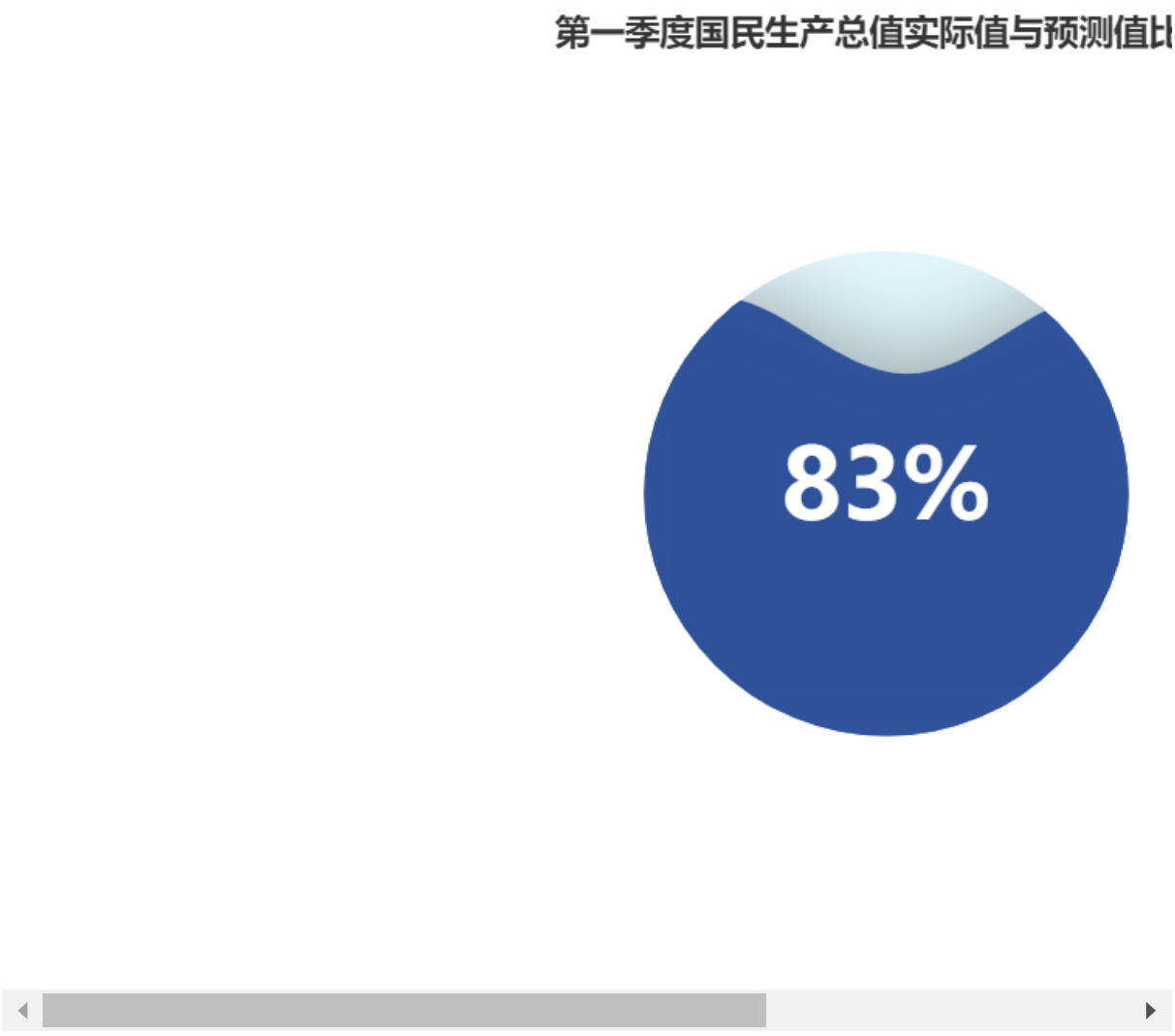
Out[20]:

```
[0.8273103019180329]
```

In [21]:

```
from pyecharts import options as opts
from pyecharts.charts import Liquid

c = (
    Liquid()
    .add("实际值/预测值", rate1, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="第一季度国民生产总值实际值与预测值比例",
                                               pos_left="center"))
)
c.render_notebook()
```

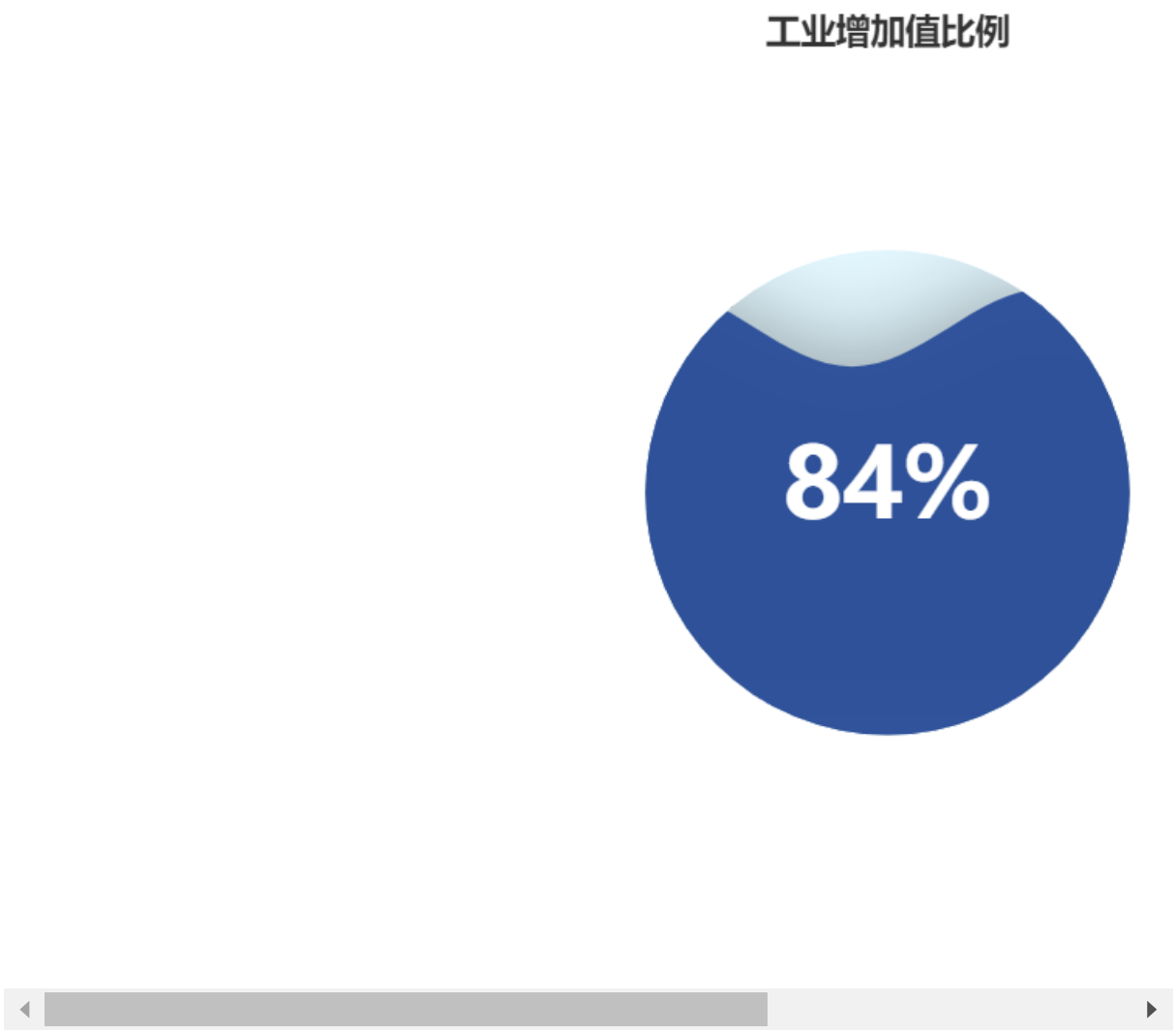
Out[21]:



In [22]:

```
warnings.filterwarnings('ignore')
data_arma = pd.DataFrame(data_economy['工业增加值'][:-2])
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()
rate2 = list(data_economy['工业增加值'][-2]/arma.forecast(steps=1)[0])
c = (
    Liquid()
    .add("实际值/预测值", rate2, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="工业增加值比例", pos_left="center"))
)
c.render_notebook()
```

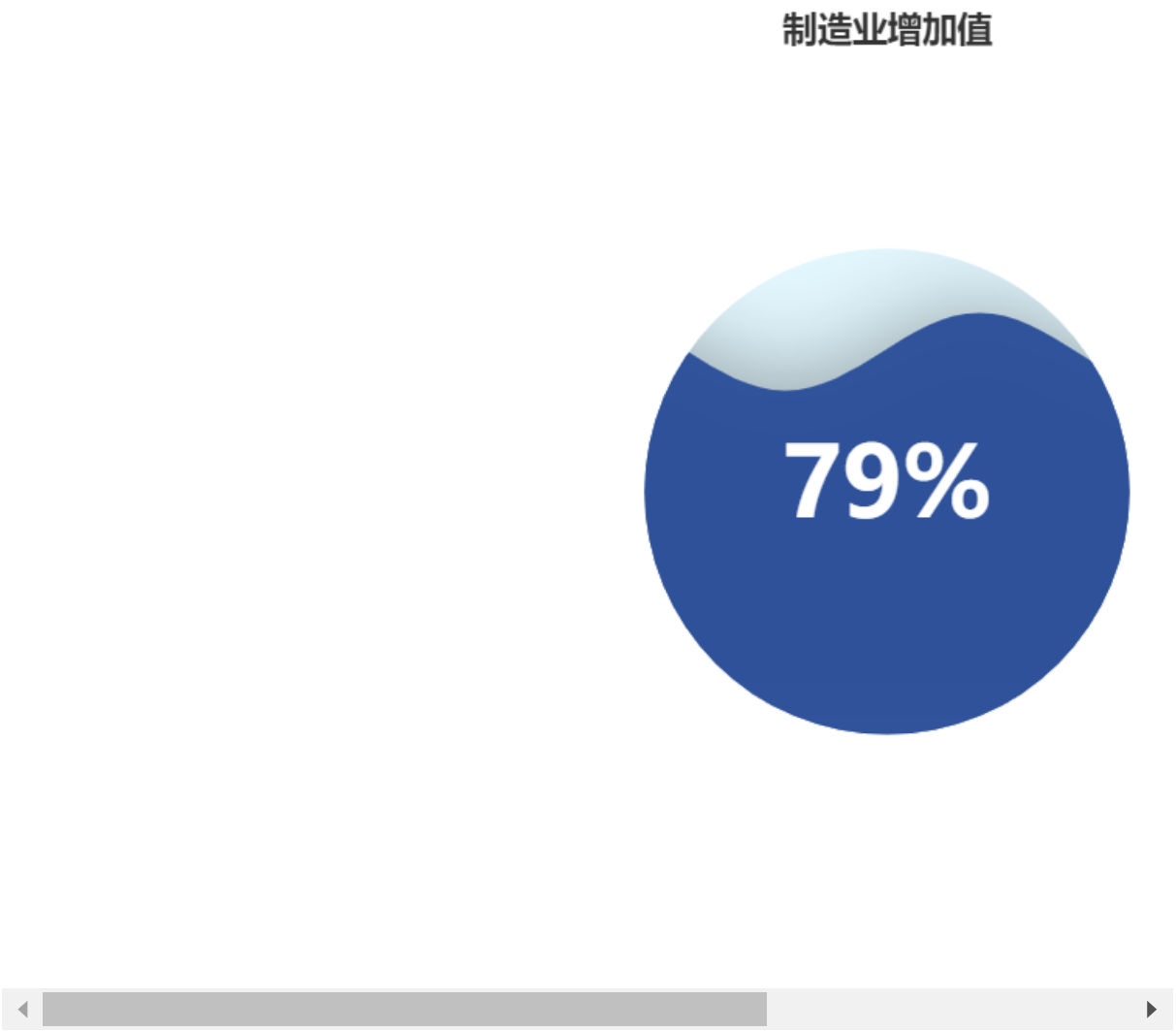
Out[22]:



In [23]:

```
warnings.filterwarnings('ignore')
data_arma = pd.DataFrame(data_economy['制造业增加值'][:-2])
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()
rate3 = list(data_economy['制造业增加值'][-2]/arma.forecast(steps=1)[0])
c = (
    Liquid()
    .add("实际值/预测值", rate3, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="制造业增加值", pos_left="center"))
)
c.render_notebook()
```

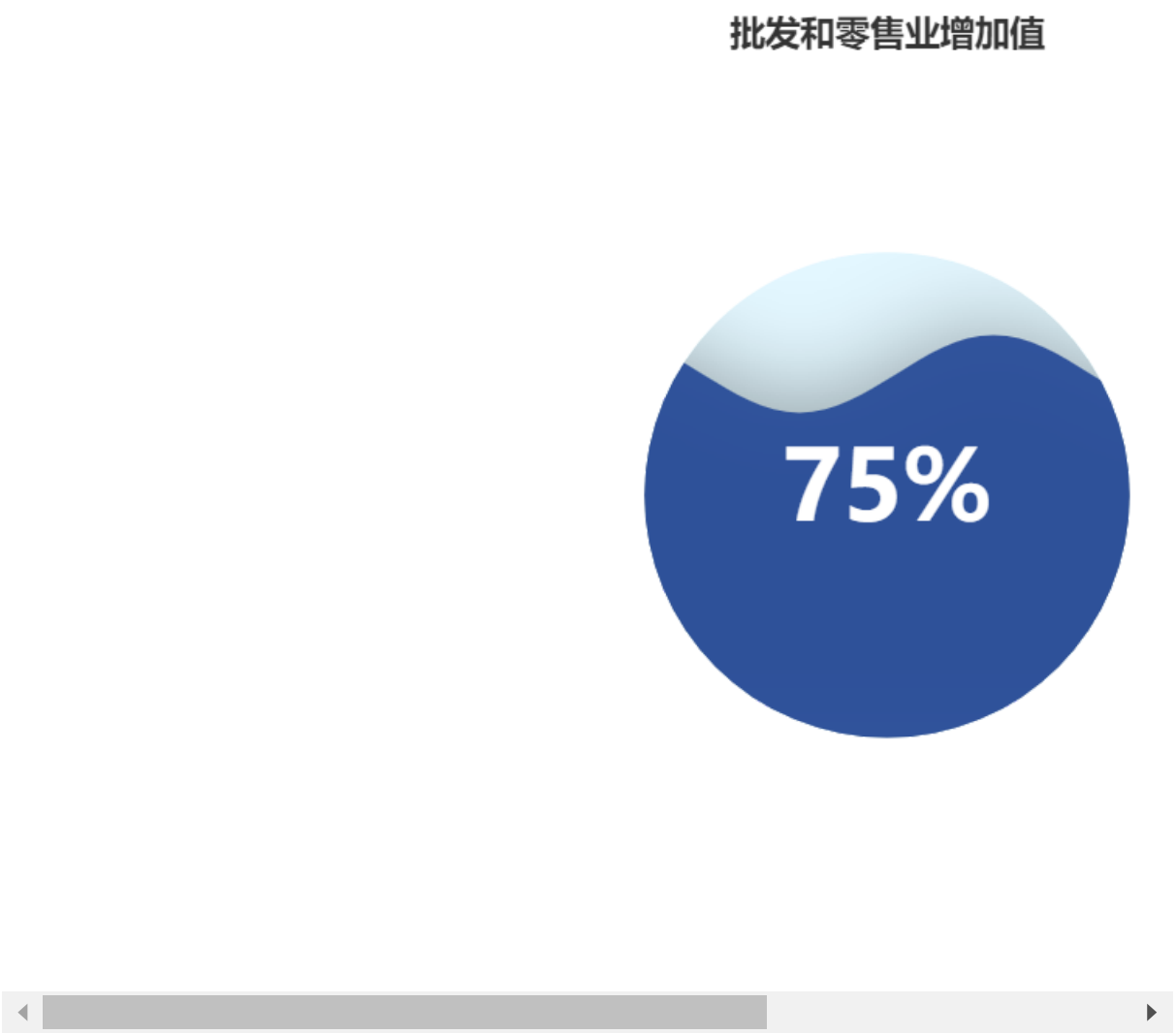
Out[23]:



In [24]:

```
data_arma = pd.DataFrame(data_economy['批发和零售业增加值'][:-2])
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()
rate4 = list(data_economy['批发和零售业增加值'][-2]/arma.forecast(steps=1)[0])
c = (
    Liquid()
    .add("实际值/预测值", rate4, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="批发和零售业增加值", pos_left="center"))
)
c.render_notebook()
```

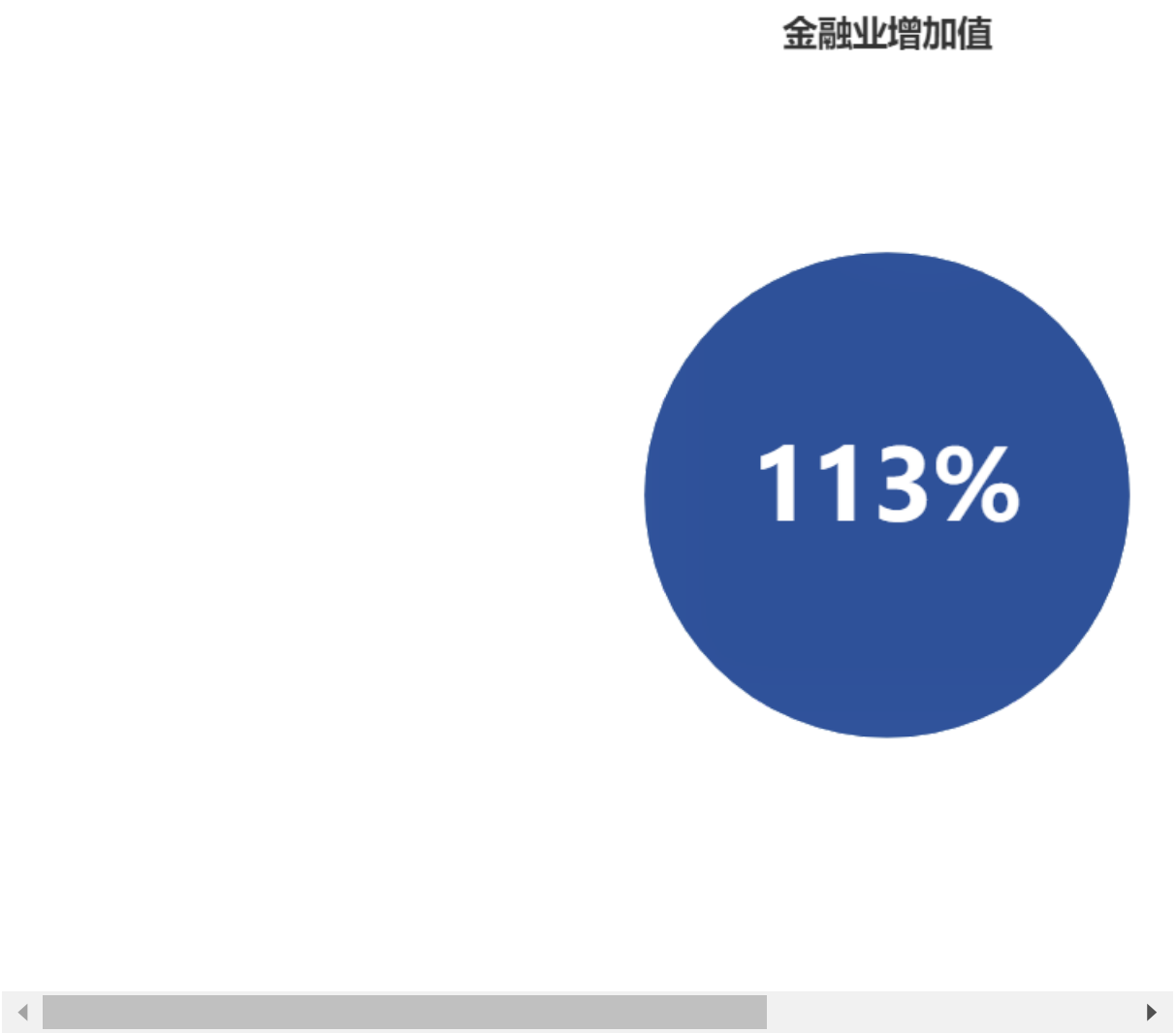
Out[24]:



In [25]:

```
data_arma = pd.DataFrame(data_economy['金融业增加值'][:-2])
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()
rate = list(data_economy['金融业增加值'][-2]/arma.forecast(steps=1)[0])
c = (
    Liquid()
    .add("实际值/预测值", rate, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="金融业增加值", pos_left="center"))
)
c.render_notebook()
```

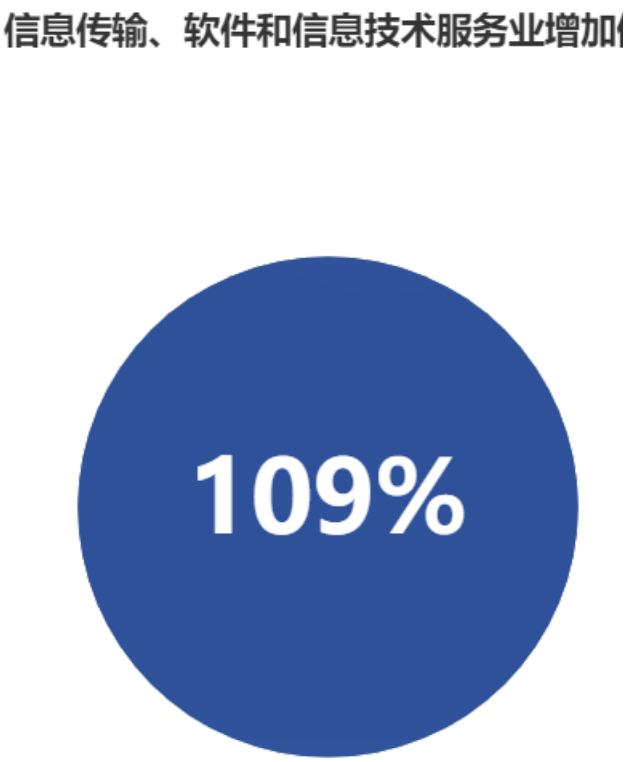
Out[25]:



In [26]:

```
data_arma = pd.DataFrame(data_economy['信息传输、软件和信息技术服务业增加值'][:-2])
a, b = arma_order_select_ic(data_arma, ic='hqic')['hqic_min_order']
arma = ARMA(data_arma, order=(a, b)).fit()
rate = list(data_economy['信息传输、软件和信息技术服务业增加值'][-2]/arma.forecast(steps=1)[0])
c = (
    Liquid()
    .add("实际值/预测值", rate, is_outline_show=False)
    .set_global_opts(title_opts=opts.TitleOpts(title="信息传输、软件和信息技术服务业增加值",
                                               pos_left="center"))
)
c.render_notebook()
```

Out[26]:



In [ ]: