期末项目评价要求

成绩构成

考核方式	总分	占总成绩的比例
课程项目及报告	100分	40%
实验项目	100分/实验，共4个实验，求取算术平均值	20%
平时作业	100分	20%
课程答辩	针对课程报告的课程答辩，100分	20%

一、成绩及说明：

课程报告及课程作品（总成绩40%）：详细要求参见二。

课程答辩（总成绩20%）：课堂答辩PPT制作及讲解，按照课程小组进行，假设20个小组，则每个小组每次讲解5分钟（100分钟/20小组），采用的大数据平台和技术架构、使用的组件和具体模块、作品设计思路、作品具体实现流程、作品实现效果分析、作品亮点和总结、确认自己实现的截屏或录像视频。

二、课程报告及课程作品具体要求和评价说明：

1、（必选）收集一份数据集：需要说明数据集的结构和格式，数据集的使用方式，需要强调如何部署到hadoop或spark平台上；

2、（必选）部署hadoop或spark平台或hadoop+spark，并能在部署好的平台上成功运行一个map-reduce算例（如wordcount等），要有部署成功和运行成功的佐证材料；

3、（必选）根据一定的应用场景或主题，基于自己收集的数据集，选取或设计一个算法（如数据挖掘的算法等，当然也可以自己写算法）并将之并行化实现，并且在部署好的大数据平台上运行成功，并结合设定好的应用场景给出算法的评价，要不同于实验；

4、（加分点，决定期末成绩差异）算法或应用的复杂性及可用性；部署且使用的平台及其模块种类（如hadoop+spark，完全分布部署；又比如利用到了Hive、图计算框架、流计算框架、MongoDB等）；算法和应用的量化可评价度量性；分析计算结果的可视化效果等；

5、（必选）需要制作课程答辩PPT，参见一中“课程答辩”要求；

6、（必选）按照提供的课程报告样例，撰写课程报告，每个小组完成一份，课程报告包括但不限于：具体的小组成员分工，平台搭建的详细过程，测试案例运行过程及结果，并行化算法设计运行的过程和结果，数据集的格式、收集、说明和使用方案，课程总结；

7、（必选）课程结束后两周内（含两周）每个小组将答辩PPT、课程报告、程序（代码+注释+运行配置说明）、数据集和媒体资源存储至百度云盘并将链接发送给助教。

》《大数据架构与技术》课程项目答辩方式

1、由于课程分组较多，分两次进行，倒数第二次课为第一次答辩，倒数第一次课为第二次答辩，已经列入本学期课程教学日历；

2、第一次答辩主要包括：数据集采集及预处理方法、系统安装部署情况、课程项目设计思路等；

3、第二次答辩主要包括：项目设计流程与算法分析、实验效果及结果评价、结果可视化展示等；

4、请大家采用PPT的排练计时功能，严格控制答辩时间，按照组号进行课程答辩；

》《大数据架构与技术》课程项目答辩方式

5、评价团队：任课教师1名、院内专家1名（按需）、企业专家1名（按需）、助教1名，对评价团队分数进行加权求和；

6、课程项目评价维度：

- 课程项目及报告，100分，期末综合成绩占比40%：数据收集及预处理15分，大数据框架使用15分，项目及算法设计25分，算法评价及结果展示25分，团队协作（含开源贡献）10分，提交资料质量10分；
- 课程答辩，100分，期末综合成绩占比20%：每次答辩50分。

2020

军事支出占GDP比重数据分析

思路

1.收集相关数据集 → 数据集上传

2.数据集数据处理 → 文本文件

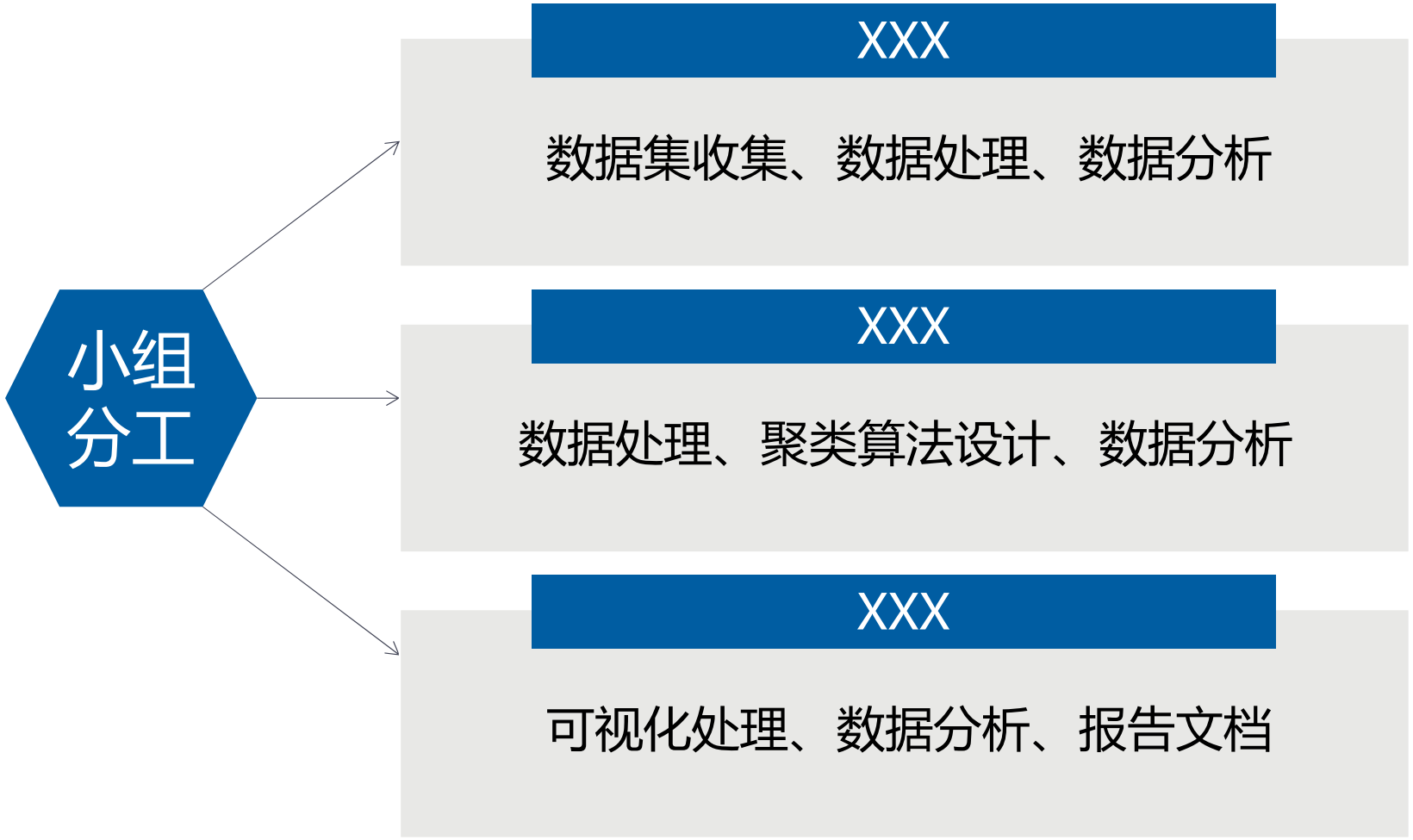
3.具体的分析算法 → 最终分析文本

目录

CONTENTS

LOGO

小组分工


小组
分工
XXX
数据集收集、数据处理、数据分析
XXX
数据处理、聚类算法设计、数据分析
XXX
可视化处理、数据分析、报告文档

目录

CONTENTS

Our World in Data

https://ourworldindata.org/military-spending

LOGO

采用数据集

military-expenditure-as-share-of-gdp

	A	B	C	D	E	F
1	Entity	Code	Year	Military expenditure (% of GDP)		
2	Afghanist	AFG	2004	2.431258		
3	Afghanist	AFG	2005	1.992068		
4	Afghanist	AFG	2006	1.896234		
5	Afghanist	AFG	2007	2.566268		
6	Afghanist	AFG	2008	2.335547		
7	Afghanist	AFG	2009	2.087413		
8	Afghanist	AFG	2010	1.945837		
9	Afghanist	AFG	2011	1.821346		
10	Afghanist	AFG	2012	1.175417		
11	Afghanist	AFG	2013	1.07695		
12	Afghanist	AFG	2014	1.298013		
13	Afghanist	AFG	2015	0.993455		
14	Afghanist	AFG	2016	0.955493		
15	Afghanist	AFG	2017	0.906857		
16	Albania	ALB	2004	1.381158		
17	Albania	ALB	2005	1.350005		
18	Albania	ALB	2006	1.567769		
19	Albania	ALB	2007	1.820765		
20	Albania	ALB	2008	1.984869		

9167	Zimbabwe	ZWE	2011	1.643824	
9168	Zimbabwe	ZWE	2012	2.263931	
9169	Zimbabwe	ZWE	2013	2.343084	
9170	Zimbabwe	ZWE	2014	2.324734	
9171	Zimbabwe	ZWE	2015	2.343629	
9172	Zimbabwe	ZWE	2016	2.198173	
9173	Zimbabwe	ZWE	2017	1.970583	
9174					

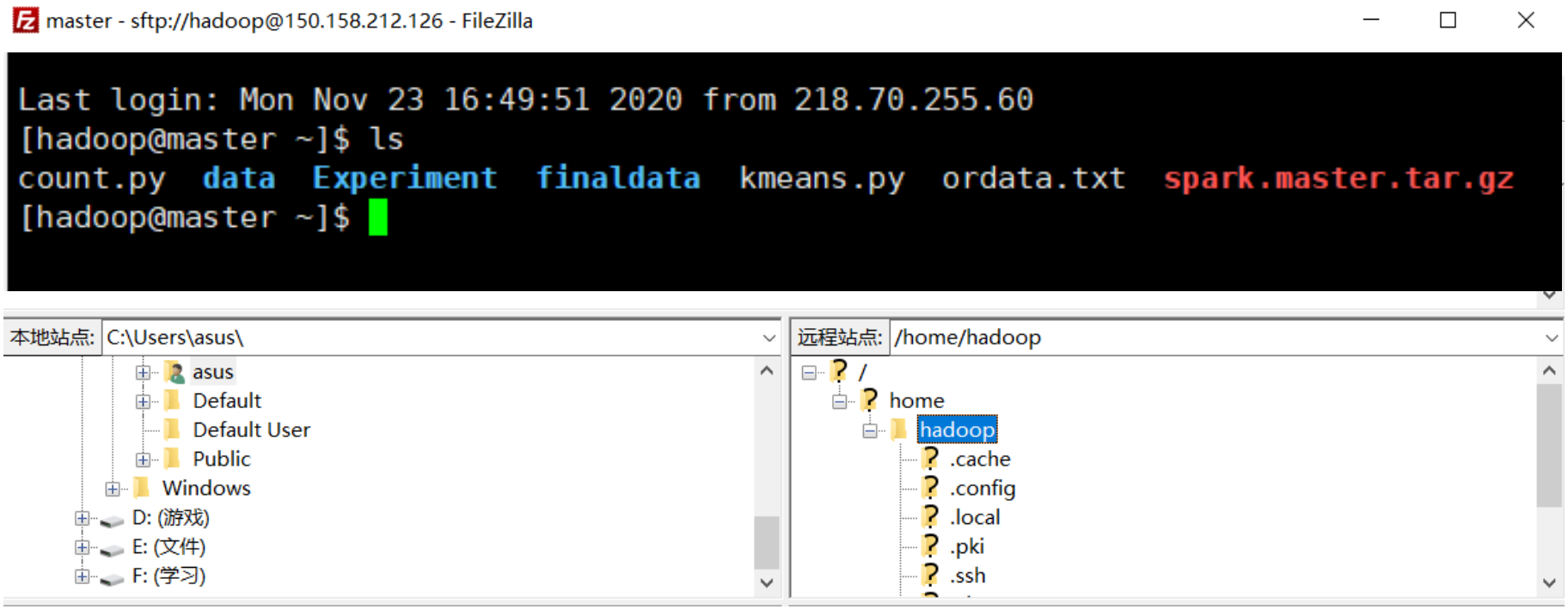
1960S-2017S

采用csv格式存储

目录

CONTENTS

FileZilla Client



数据第一次筛选

采用2004S-2017S的数据

Hdfs： 数据第二次筛选

```
from
from
impor
conf = Spark(
sc = Sp
client
```

```
lines = [(row.split(',')[0],float(row.split(',')[3]))for row in lines]
data = sc.parallelize(lines)
data_1 = data.groupByKey()
data_2 = data_1.mapValues(lambda a:[i for i in a])
```

```
08235108]), ('Angola', [3.471180187, 3.692240947, 3.761495305, 3.114054493, 3.57312
```

```
data_2.saveAsTextFile("file:///home/hadoop/data")
```

```
1984, 5.840526749, 6.389929368, 7.506153829, 6.202371753, 6.410466315]), ('Argentin
a', [0.883295359, 0.847154784, 0.788233877, 0.792548175, 0.762756475, 0.886508577,
0.814878105, 0.764287094, 0.784824721, 0.83773643, 0.878100919, 0.86472284, 0.82713
0106, 0.909662993]), ('Armenia', [2.741996705, 2.871931363, 2.947613156, 3.04257470
1, 3.395579363, 4.157049908, 4.265646056, 3.853946441, 3.583766499, 3.997244557, 3.
943378265, 4.239225803, 4.080226863, 3.968039367]), ('Australia', [1.827410769, 1.8
01853775, 1.821842846, 1.814599669, 1.797239012, 1.929179691, 1.862399774, 1.769389
864, 1.679917358, 1.649533076, 1.781294955, 1.958451254, 2.092849524, 1.989292632])
, ('Austria', [0.893480466, 0.853724571, 0.789783772, 0.90565864, 0.876168945, 0.83
8784296, 0.824768861, 0.794738036, 0.782234948, 0.754017344, 0.753954082, 0.7070986
42, 0.747000092, 0.729355232]), ('Azerbaijan', [2.629481138, 2.299069675, 3.4177593
```

目录

CONTENTS

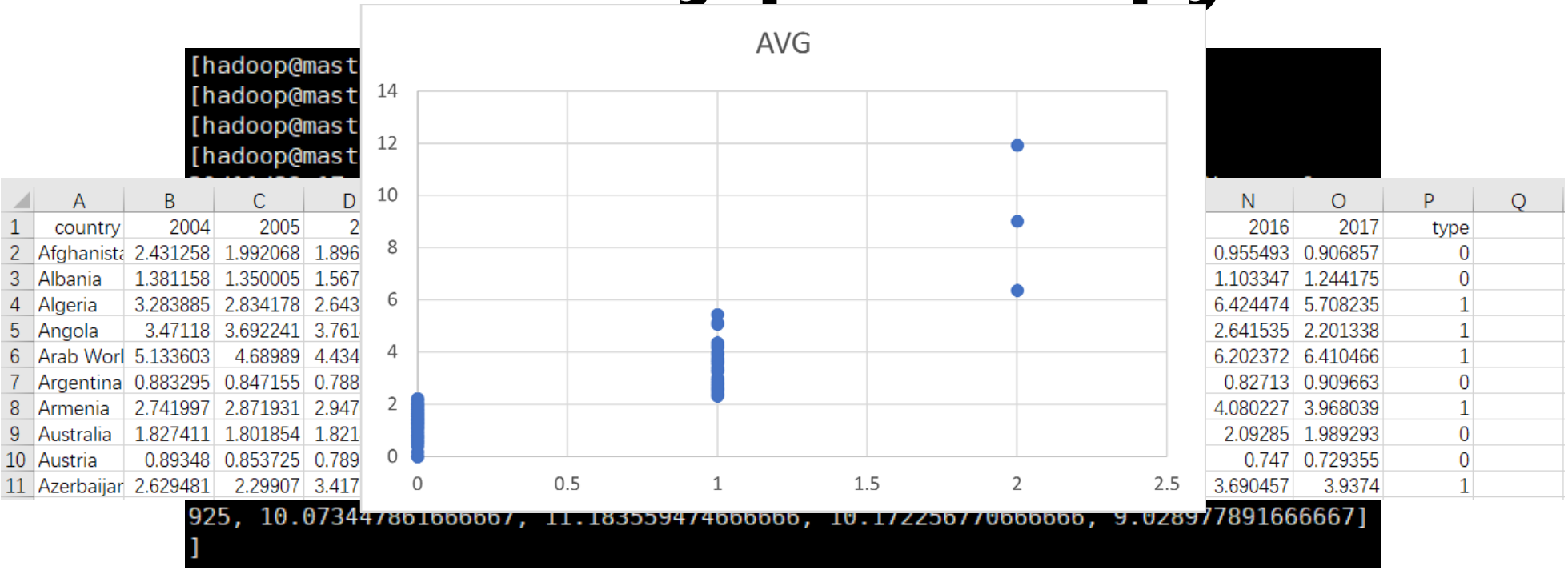
调用SPARK库 数据过滤

```
K = int(3)
convergeDist = float(0.1)

kPoints = data.takeSample(False, K, 1)
tempDist = 2.0

while tempDist > convergeDist:
    closest = data.map(
        lambda p: (closestPoint(p, kPoints), (p, 1)))
    pointStats = closest.reduceByKey(
        lambda p1_c1, p2_c2: ([i[0]+i[1] for i in zip(p1_c1[0],p2_c2[0])], p1_c
1[1] + p2_c2[1]))
    newPoints = pointStats.map(
        lambda st: (st[0], [i/st[1][1] for i in st[1][0]])).collect()
    tempDist = sum(np.sum([(i[0]-i[1])**2 for i in zip(kPoints[iK],p)]) for (iK
, p) in newPoints)
    for (iK, p) in newPoints:
        kPoints[iK] = p
print("Final centers: " + str(kPoints))
closest = lines.map(lambda p: (p[0],p[1],closestPoint(p[1], kPoints)))
closest.saveAsTextFile("file:///home/hadoop/finaldata")
```

LOGO

数据集处理

fidataf by a pdrotaess.py

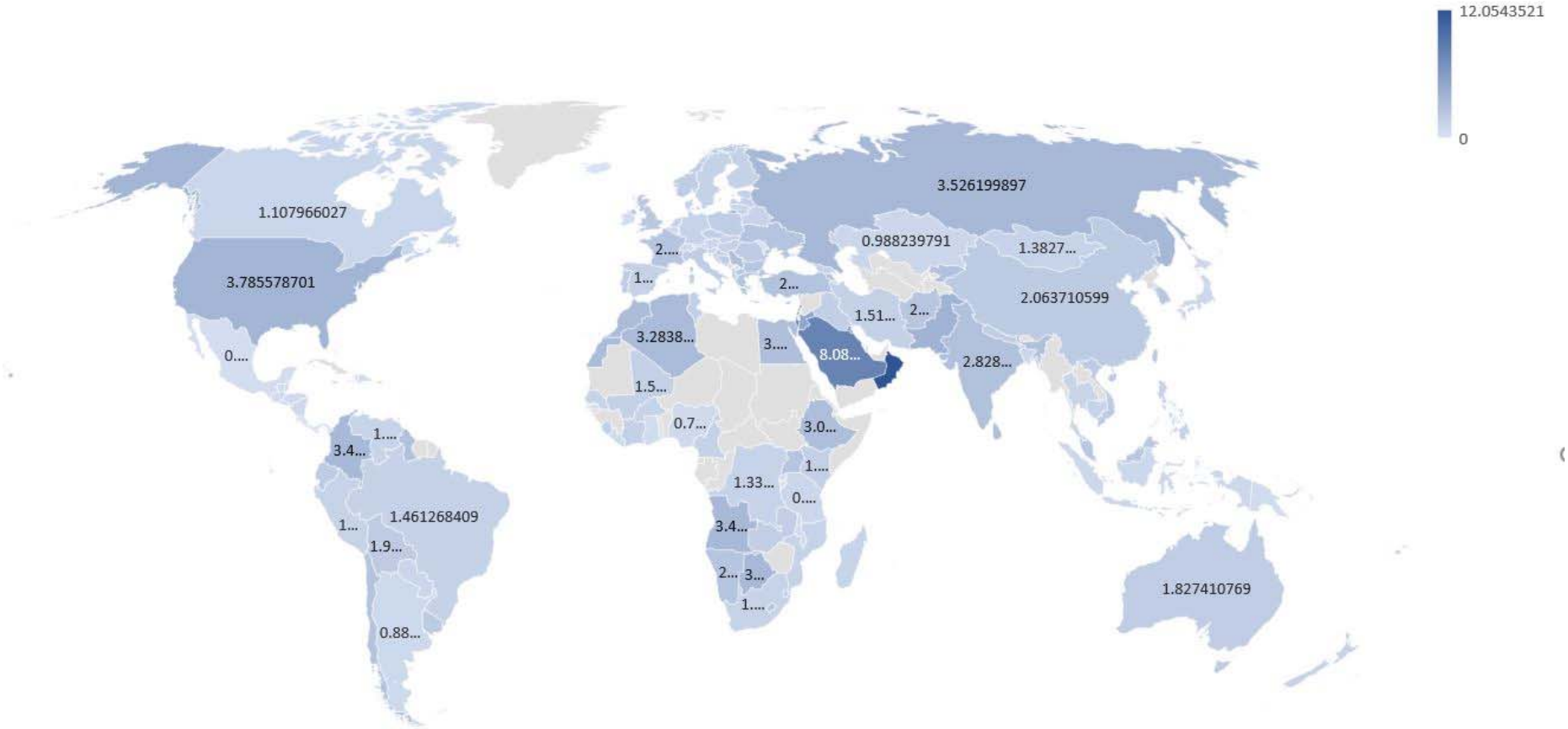


	A	B	C	D	N	O	P	Q
1	country	2004	2005	2	2016	2017	type	
2	Afghanist	2.431258	1.992068	1.896	0.955493	0.906857	0	
3	Albania	1.381158	1.350005	1.567	1.103347	1.244175	0	
4	Algeria	3.283885	2.834178	2.643	6.424474	5.708235	1	
5	Angola	3.47118	3.692241	3.761	2.641535	2.201338	1	
6	Arab Worl	5.133603	4.68989	4.434	6.202372	6.410466	1	
7	Argentina	0.883295	0.847155	0.788	0.82713	0.909663	0	
8	Armenia	2.741997	2.871931	2.947	4.080227	3.968039	1	
9	Australia	1.827411	1.801854	1.821	2.09285	1.989293	0	
10	Austria	0.89348	0.853725	0.789	0.747	0.729355	0	
11	Azerbaijar	2.629481	2.29907	3.417	3.690457	3.9374	1	

目录

CONTENTS

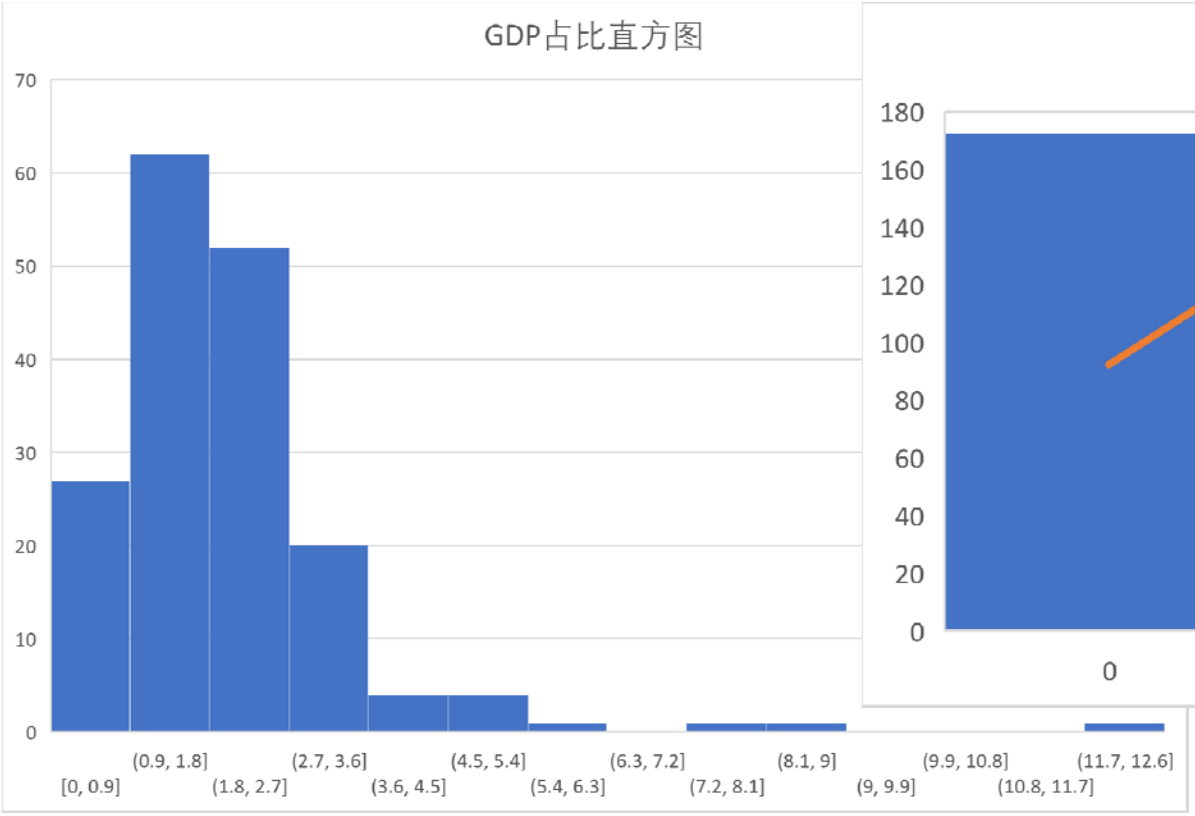
数据地图

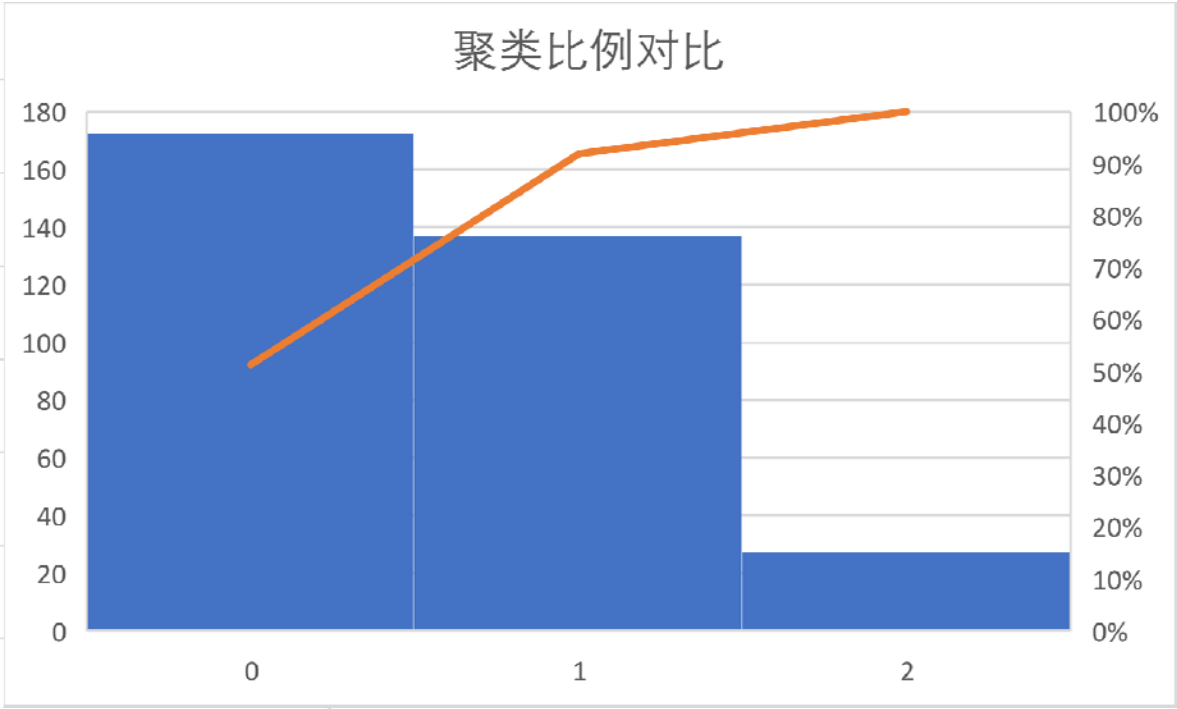

12.0543521
0
1.107966027
3.785578701
0....
1....
3.4...
1...
1.461268409
1.9...
0.88...
3.526199897
0.988239791
1.3827...
2....
1...
2...
2.063710599
1.51...
2...
3.2838...
3....
8.08...
2.828...
1.5...
0.7...
3.0...
1....
1.33...
0....
3.4...
2...
3...
1....
1.827410769

LOGO

结果分析

数据占比

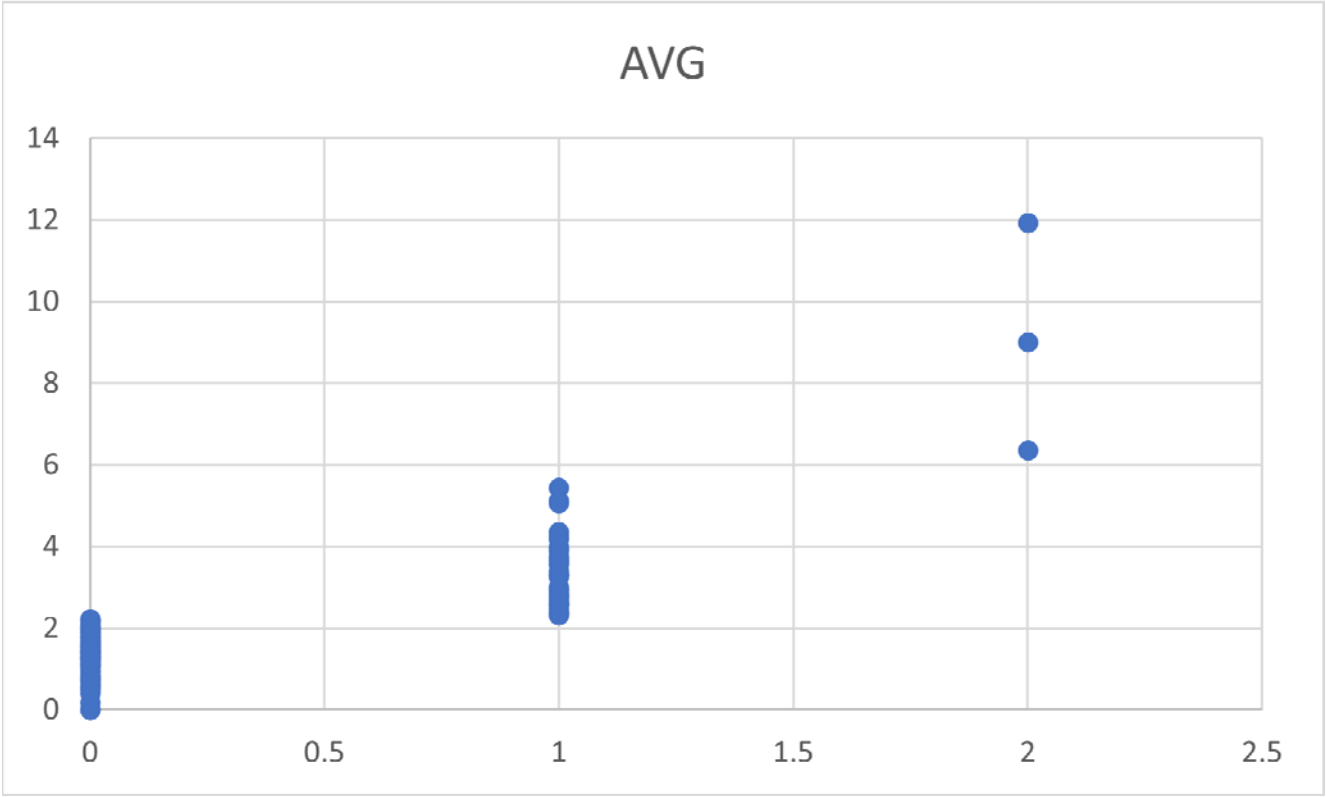




LOGO

结果分析

聚类散点图



聚3类

LOGO

结果分析

第二类

78	Israel	8.185524	7.637096	7.4	.670257	4.654133	4.725543	2
126	Oman	12.05435	11.81698	10.8	14.3842	15.95615	12.07272	2
140	Saudi Arabia	8.08156	7.730601		13.49623	9.906483	10.28867	2

第一类

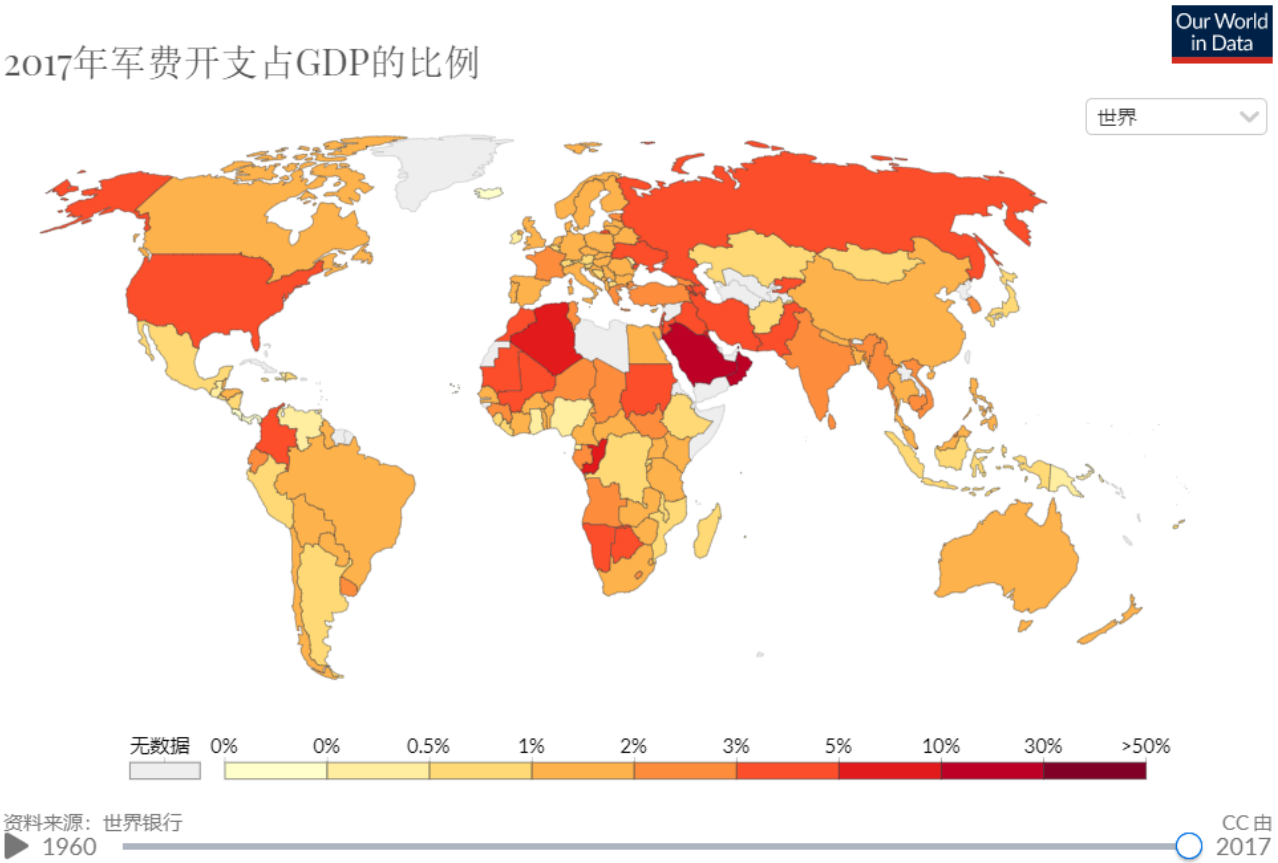
1	2
美国 3.902	伊朗 2.591
俄罗斯 3.937	巴基斯坦 3.394
	印度 2.582

第零类

1数量众多

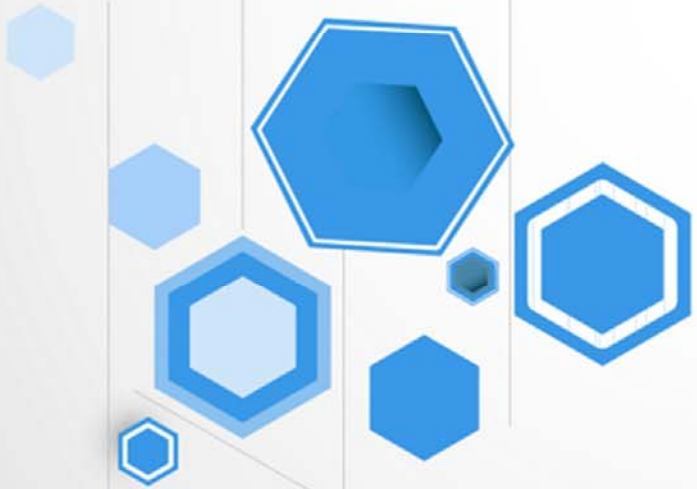
2分布广泛

3以北欧、西欧、
泛亚、南非国家为主

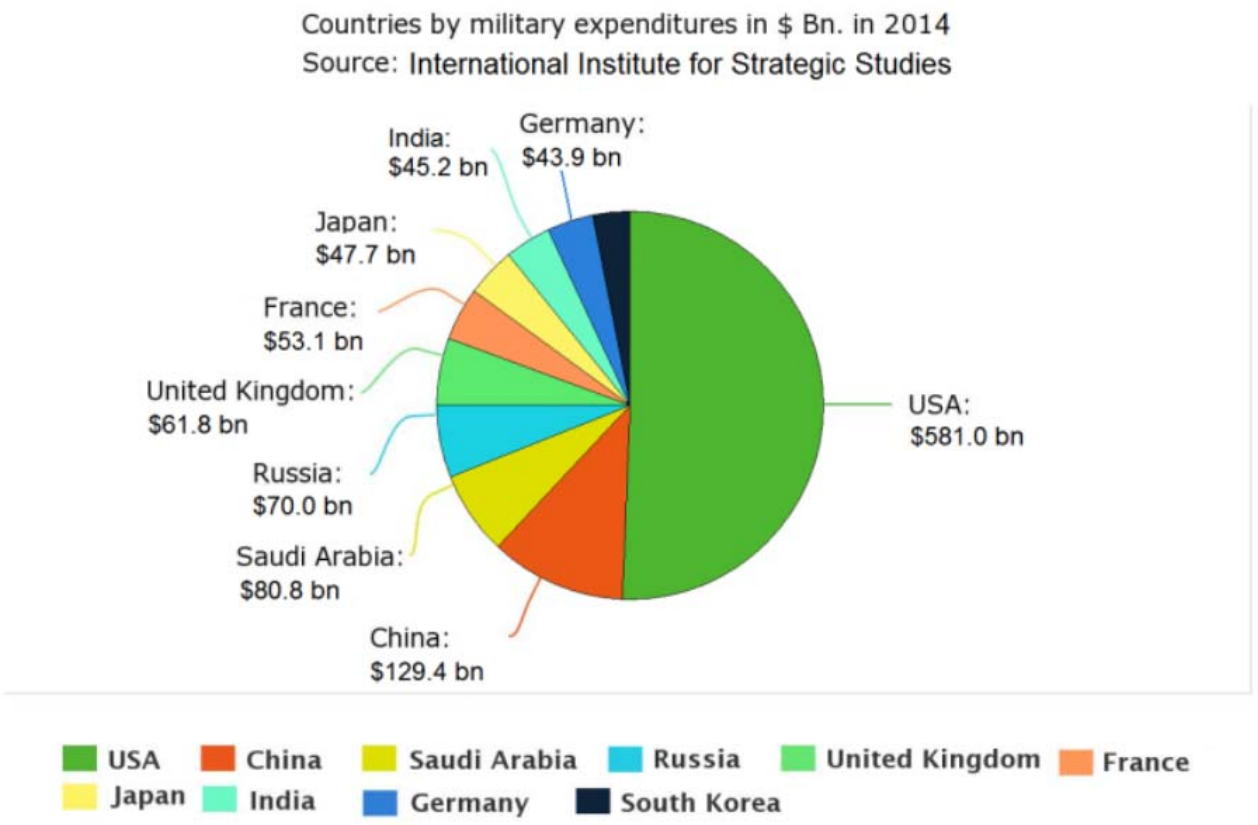


目录

CONTENTS



缺点


Countries by military expenditures in $ Bn. in 2014
Source: International Institute for Strategic Studies
India: $45.2 bn
Germany: $43.9 bn
Japan: $47.7 bn
France: $53.1 bn
United Kingdom: $61.8 bn
Russia: $70.0 bn
Saudi Arabia: $80.8 bn
China: $129.4 bn
USA: $581.0 bn
USA
China
Saudi Arabia
Russia
United Kingdom
France
Japan
India
Germany
South Korea

期末项目评价要求

Thank You!