大数据引论开题答辩

THE世界大学排名数据分析

基于2011-2024的泰晤士高等教育世界大学排名的数据分析

数据集来源：https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings-2011-2023

https://www.kaggle.com/datasets/ddosad/ti- mesworlduniversityrankings2024

刘琮璟　　徐朱玮

目录 CATALOGUE

01 研究方向和预期目标
RESEARCH DIRECTION AND EXPECTED OBJECTIVES

02 数据来源和介绍
INTRODUCTION AND RESOURCE OF DATASET

03 研究任务和方法
RESEARCH TARGETS AND METHODS

04 预期成果和时间安排
EXPECTED RESULTS AND SCHEDULE

01

研究方向和预期目标

RESEARCH DIRECTION AND EXPECTED OBJECTIVES

01 研究方向和预期目标——选题背景

RESEARCH DIRECTION AND EXPECTED OBJECTIVES

THE

对于留学生而言，最值得关心的事情之一就是院校排名。

由英国《泰晤士高等教育》（Times Higher Education，简称THE）发布的泰晤士高等教育世界大学排名（Times Higher Education World University Rankings），就是各大机构发表的排名中最权威的之一。

据最新版《全国研究生招生调查报告》显示，我国今年考研人数再创新高，达到了474万人次。在学历膨胀、各行各业内卷严重的大环境下，许多高校的毕业生都选择避开人山人海的“国内升学之路”，选择去海外攻读研究生，提升学历背景。

据教育部数据统计，2017年，我国出国留学人数首次突破60万大关，达到60.84万人；2018年度我国出国留学人员总数升至66.21万人；2019年度我国出国留学人员总数为70.35万人。三年反反复复的疫情使得大部分留学党只能转战国内，但随着疫情结束，预计2023年的留学人数将突破80万。

选题的意义

相比于国内更注重的QS排名和申请美国更注重的U.S.News，以及代表了中国大学评价话语体系的软科排名，THE大学排名似乎比较少被大家提起。但事实上，THE排名考察了教学、研究环境、研究质量、产业和国际声望5个大类下的18个绩效指标。也被民间誉为是在学术评价上最权威的一项排名。

充分分析和了解THE大学排名，一方面能够帮助同学们在选校的地域、国家还有学校研究能力等指标上有更多的参考；另一方面也能够帮助国内头部的高校分析自身与排名中更高水平的学校的差距，从而找到自身的特色和提升的方向。

- 增进对THE排名的了解
- 帮助留学生增加择校参考指标
- 帮助国内高校找到提升参考

01 研究方向和预期目标——预期目标

RESEARCH DIRECTION AND EXPECTED OBJECTIVES

- 最终整个项目将被呈现为网页形式，以下四个部分都将成为网页的组成

- 可视化

由于世界大学排名官网都仅有排名数据而没有视图，因此将使用已有的THE数据集完成一系列数据的可视化，让用户有直观的感受

- 自主调权重

所有的世界大学排名都有自己的评价体系，但是每个评价指标的权重也是固定的

我们希望能够让用户需自主调节指标权重，生成个性化排名

- 2025排名预测

利用过去十年已有的数据，我们希望有可能对2025年的THE大学排名进行预测

给计划将来出国的学生提供参考，也给相关大学提供指标提升参考

- 对国内高校的建议

我们将着重分析清华、北大、复旦、交大、浙大，这5所大学的相关数据，并将之与其它高排名院校进行对比

以期能够为国内高校的发展提供帮助

02

数据来源和介绍

INTRODUCTION AND RESOURCE OF DATASET

数据来源

本研究所用的数据集来源于Kaggle，这是一个为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。上面有很多数据集可供下载，并用于分析训练。

我们在上面获得了从2011-2023年的THE世界大学排名数据（https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings-2011-2023）和2023年的THE世界大学排名最新数据（https://www.kaggle.com/datasets/ddosad/timesworlduniversityrankings2024）。据Kaggle上的相关介绍，这两份数据集都是从https://www.timeshighereducation.com/world-university-rankings官网上直接获得的。

02 数据来源和介绍——数据介绍

INTRODUCTION AND RESOURCE OF DATASET

这份数据集包含了从2011-2024年THE世界大学排名的全部数据。

其中每一年所含的参与排名大学数据量不同，具体为：

2011：199　　2012：402

2013：400　　2014：400

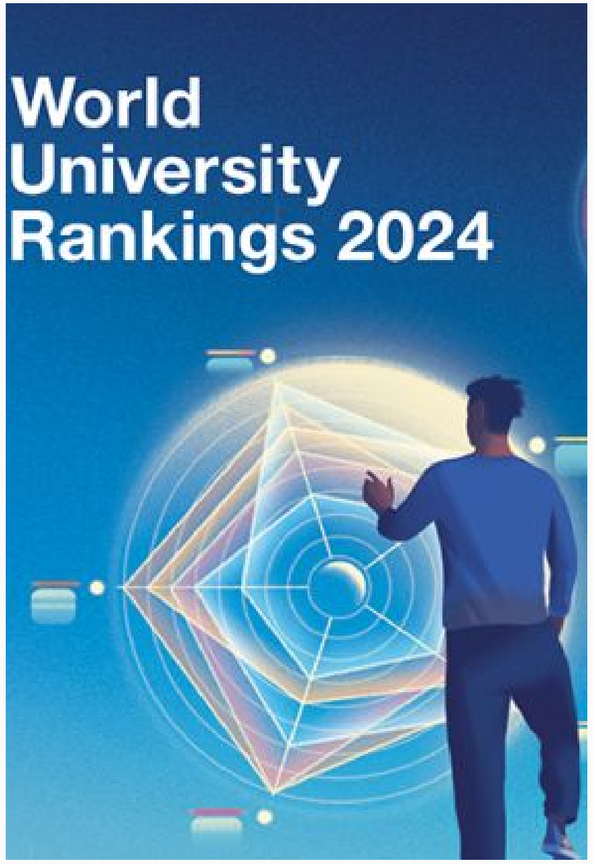
2015：401　　2016：800

2017：981　　2018：1103

2019：1258　2020：1397

2021：1526　2022：2112

2023：2345　2024：2673



此外，每一年的评价指标也有所不同。在2016年和2024年，分别发生了评价指标的增多。

2011-2015年的评价指标为：教学、国际声誉、行业收入、研究、引用、国家、学校名、专业；2016-2023年增加了：学生数、师生比、国际学生占比、男女比例；2024年则又增加了学校类别（公立私立等）和学校资格级别等。

样本量和特征都随着年份增长总体增加。

02 数据来源和介绍——数据示例

INTRODUCTION AND RESOURCE OF DATASET

	0	1	2	3	4	5
rank_order	1	2	3	4	5	6
rank	1	2	3	4	5	6
name	Harvard University	California Institute of Technology	Massachusetts Institute of Technology	Stanford University	Princeton University	University of Oxford
scores_overall	96.1	96.0	95.6	94.3	94.2	91.2
scores_overall_rank	1	2	3	4	5	7
scores_teaching	99.7	97.7	97.8	98.3	90.9	88.2
scores_teaching_rank	1	4	3	2	6	9
scores_international_outlook	72.4	54.6	82.3	29.5	70.3	77.2
scores_international_outlook_rank	49	93	36	156	53	42
scores_industry_income	34.5	83.7	87.5	64.3	-	73.5
scores_industry_income_rank	105	24	21	33	164	28
scores_research	98.7	98.0	91.4	98.1	95.4	93.9
scores_research_rank	2	4	11	3	5	8
scores_citations	98.8	99.9	99.9	99.2	99.9	95.1
scores_citations_rank	8	1	2	6	3	22
location	United States	United States	United States	United States	United States	United Kingdom
aliases	Harvard University	California Institute of Technology caltech	Massachusetts Institute of Technology	Stanford University	Princeton University	University of Oxford
subjects_offered	Mathematics & Statistics,Civil Engineering,Lan...	Languages, Literature & Linguistics,Economics ...	Mathematics & Statistics,Languages, Literature...	Physics & Astronomy,Computer Science,Politics ...	Languages, Literature & Linguistics,Biological...	Accounting & Finance,General Engineering,Commu...
closed	False	False	False	False	False	False
unaccredited	False	False	False	False	False	False

这是2011年前6所大学的数据示例，将原本的csv文件导入python，并且处理成了DataFrame的格式。由于数据特征过多、长度过长，因此这里做了转置以便呈现。

可以看到，数据特征中有每所学校对应的教学、国际声誉、行业收入、研究、引用、国家、学校名、专业等指标，可以用于可视化和数据分析。

03

研究任务和方法

RESEARCH TARGETS AND METHODS

主要的任务分为四个部分：

1、数据可视化

2、改变权重生成新的大学排名

3、分析预测2025年THE排名

4、个性化分析清北交复浙5所高校

03 研究任务和方法——数据可视化

RESEARCH TARGETS AND METHODS

当数据作为表格存在时更容易查找，但不够直观。而几乎所有的世界大学排名官网都只有表格数据而没有视图。因此，我们希望能够将THE世界大学排名（2011-2024）中最有效和需要被直观了解的部分可视化。

由于一共有14年的数据，因此这一可视化任务的思路主要分为两大类：第一类，展现从2011-2024年的数据走势；第二类，展现特定年份的特征雷达图。此外，还将按照“国家”、“学校”和“排名”进行分类。

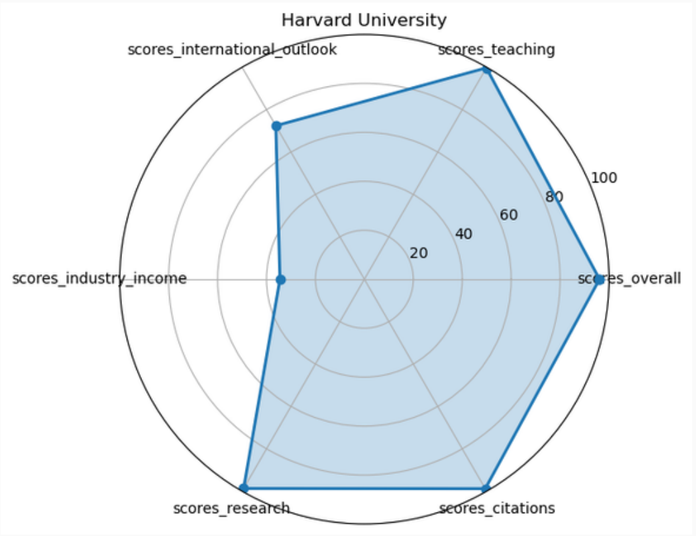
预计的可视化视图有：

1）每一年不同国家上榜的大学的数量；

2）每个国家近十年来上榜大学数量的变化；

3）按照国家、学校、排名，对每个指标都做变化图；

4）按照国家、学校、排名，给出国家均值的各项指标雷达图。



示例：采用2011年哈佛大学数据所画的学校各项指标雷达图

可视化方法

这里的可视化任务主要将通过python中的matplotlib库完成，辅以pandas和numpy。

预计主要用到的视图类型有折线图、饼图、柱状图、雷达图。如果后续发现更好的可视化方案也会再做提升。

03 研究任务和方法——重排名

RESEARCH TARGETS AND METHODS

THE世界大学排名主要依据五个指标，其中包括教学质量（教学环境）30%，研究(论文发表数量、收入和声誉)30%，引文(研究影响)30%，国际展望(工作人员、学生和研究)7.5%，和产业收入(知识转化)2.5%。

预期效果

由于在不同指标下排名变动较大，而每个人关注的指标不同，我们希望可以根据不同的需求，制定个性化排名。基于2024年最新数据，当用户输入不同指标所占权重时，可以直接输出定制化全新排名。

方法：pandas和numpy

rank	name
1	University of Oxford
2	Stanford University
3	Massachusetts Institute of Technology
4	Harvard University
5	University of Cambridge
6	Princeton University
7	California Institute of Technology
8	Imperial College London
9	University of California, Berkeley
10	Yale University

示例： THE原综合排名

rank	name
1	University of Cambridge
2	University of Oxford
3	Harvard University
4	University of California, Berkeley
5	Tsinghua University
6	California Institute of Technology
7	Princeton University
8	Stanford University
9	Peking University
10	ETH Zurich

研究情况排名（研究100%）

03 研究任务和方法——分析预测2025年THE排名

RESEARCH TARGETS AND METHODS

	rank	name	scores_overall	scores_teaching	scores_research	scores_citations	scores_industry_income	scores_international_outlook
0	1	Harvard University	96.1	99.7	98.7	98.8	34.5	72.4
1	2	Harvard University	93.9	95.8	97.4	99.8	35.9	67.5
3	4	Harvard University	93.6	94.9	98.6	99.2	39.9	63.7
1	2	Harvard University	93.9	95.3	98.5	99.1	40.6	66.2
1	2	Harvard University	93.3	92.9	98.6	98.9	44.0	67.6
5	6	Harvard University	91.6	83.6	99.0	99.8	45.2	77.2
5	6	Harvard University	92.7	87.5	98.3	99.7	47.3	77.9
5	6	Harvard University	91.8	84.2	98.4	99.7	46.4	79.7
5	6	Harvard University	93.6	90.1	98.4	99.6	48.7	79.7
6	7	Harvard University	93.0	89.2	98.6	99.1	47.3	76.3
2	3	Harvard University	94.8	94.4	98.8	99.4	46.8	77.7
2	=2	Harvard University	95.0	94.5	98.9	99.2	48.9	79.8
1	2	Harvard University	95.2	94.8	99.0	99.3	49.5	80.5
3	4	Harvard University	97.8	97.7	99.9	99.4	84.2	90.8

X_train

y_train

X_test

y_test

由于我们没有2025年的各项数据指标，因此预测下一年的排名变得非常困难。

目前我们初步的设想是，通过错位的方法划分训练集和测试集，也即：比如将2011年的各项指标对应到2012年的综合评分，通过这样的错位方式来用2024年已有的数据预测2025年的各大高校综合评分，以此对2025年的高校排名进行预测。

这一想法基于每所高校相邻两年间的排名一般不会发生巨大的变化，但是在前期实验的准确性上表现仍然不佳。希望后续可以找到更合适的预测方法。

03 研究任务和方法——国内top5大学个性化分析

RESEARCH TARGETS AND METHODS

聚焦国内

从2011年到2024年，中国整体教育的迅速发展带来了中国大学排名的显著上升。在此期间，清华大学、北京大学、复旦大学、上海交通大学和浙江大学这五所知名学府的表现尤为突出。

“领头羊”：2011 北京大学 37 ------ 2024 清华大学 12

“黑马”：2011 浙江大学 197 ------ 2024 浙江大学 55

研究方法

使用python进行可视化数据分析，将数据与其他资料（如大学官网信息、媒体公众号报道等） 相结合，针对不同大学的不同特点进行个性化分析，为高校后续发展及学生择校提供帮助。

预期效果

1、使用折线图，分析在2011年-2024年期间每所高校排名走势变化，判断整体趋势。

2、结合五项评价指标与单项排名（教学质量、研究质量、引文情况、国际展望、产业收入），对比对比顶尖高校，分析每所大学的强势方面和弱势方面，给出合理化提升建议。

3、基于五所学校的基础指标，如师生比例、男女比例、国际生比例、学科覆盖情况等，分析可提升空间，给出后续发展建议。

04

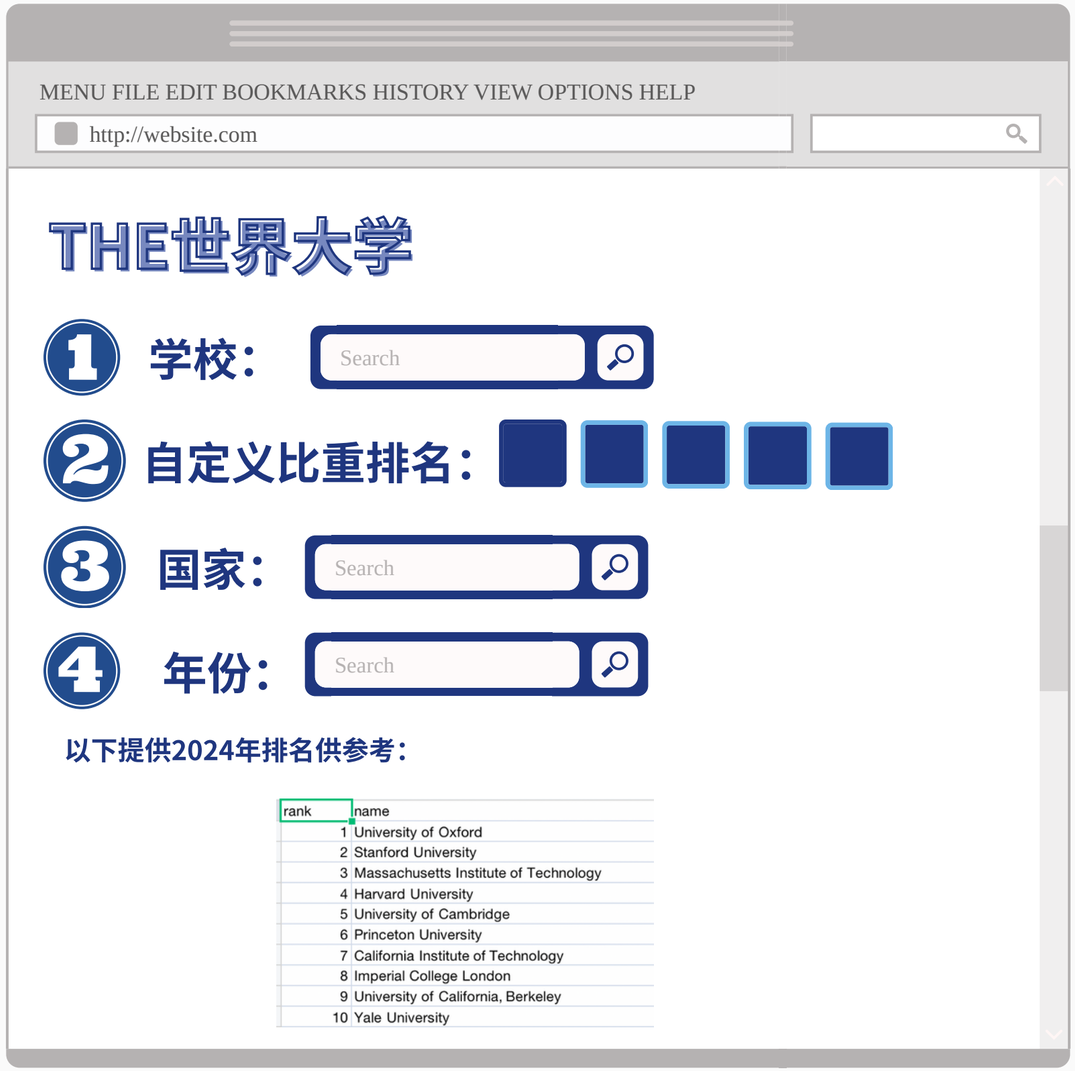
预期成果和时间安排

EXPECTED RESULTS AND SCHEDULE

04 预期成果和时间安排

EXPECTED RESULTS AND SCHEDULE

预期网页效果



由于信息冗杂，数据和图表过多，我们希望制作可交互式网页，帮助用户更加方便快捷地找到所需信息。

1. **学校：输入学校名称，输出该学校相关信息（如最新指标雷达图和近十年排名走势）**
2. **自定义：输入自定义各项指标权重，输出全新排名**
3. **国家：输入国家名称，输出有关该国家的信息（如最新上榜大学名单、各指标均值雷达图）**
4. **年份：输入年份，输出当年排行榜（包括25年预测排行）**

04 预期成果和时间安排

EXPECTED RESULTS AND SCHEDULE

第8周（10月24日）：开题

查找数据集-确定选题-研究背景意义-讨论预期目标及呈现效果-判断研究方法及其可行性-正式开题

第11周（11月中旬）

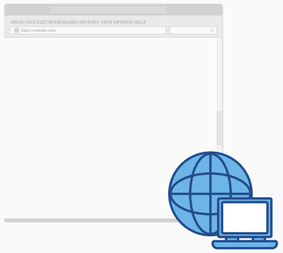
完成基础的数据处理及模型搭建工作：

（1）完成所有可视化图表（包括折线图、柱状图、雷达图等）的制作；

（2）完成“更改评价指标比重再排名”模型的搭建；

（3）对比选择合适的预测模型和方法，完成框架搭建；

（4）确定五所大学个性化分析的具体内容，找出其中具有研究意义的重点内容。



第14周（12月上旬）

（1）完成html建设，将python代码与网站进行前后端连接，对网页进行调试并完善；

（2）对排名预测模型进行优化，输出结果；

（3）结合数据和资料信息，完成对五所大学的个性化分析，落实为书面文字。

（4）整体回顾研究内容及实验结果，找出不足之处并完善。

第16周（12月下旬）：

完成最终课程论文撰写并提交报告。

请老师和同学们批评指正

THANK THE TEACHER AND STUDENTS FOR THEIR CRITICISM AND CORRECTION

徐朱玮

刘琮璟