# 第五次实验：全球高校学科数据分析与建模

## 一、实验目的

本实验旨在通过对全球高校 ESI 学科排名数据的综合分析，探索高校的学科分布特征、聚类规律以及华东师范大学（ECNU）的学科画像，并基于机器学习模型实现学科排名预测。

## 二、数据来源与预处理

数据来自 Clarivate ESI（Essential Science Indicators）数据库，各学科 CSV 文件包含：学校名称、国家/地区、Web of Science 论文数（Docs）、被引次数（Cites）、平均被引（Cites/Paper）、高被引论文数（Top Papers）。

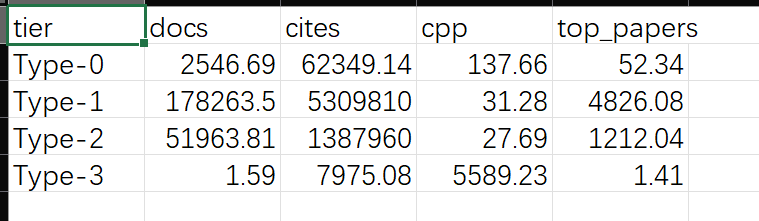
使用 merge\_subjects.py 对所有学科文件进行合并，跳过首行说明，统一重命名为标准字段：['univ', 'country', 'docs', 'cites', 'cpp', 'top\_papers', 'subject']。

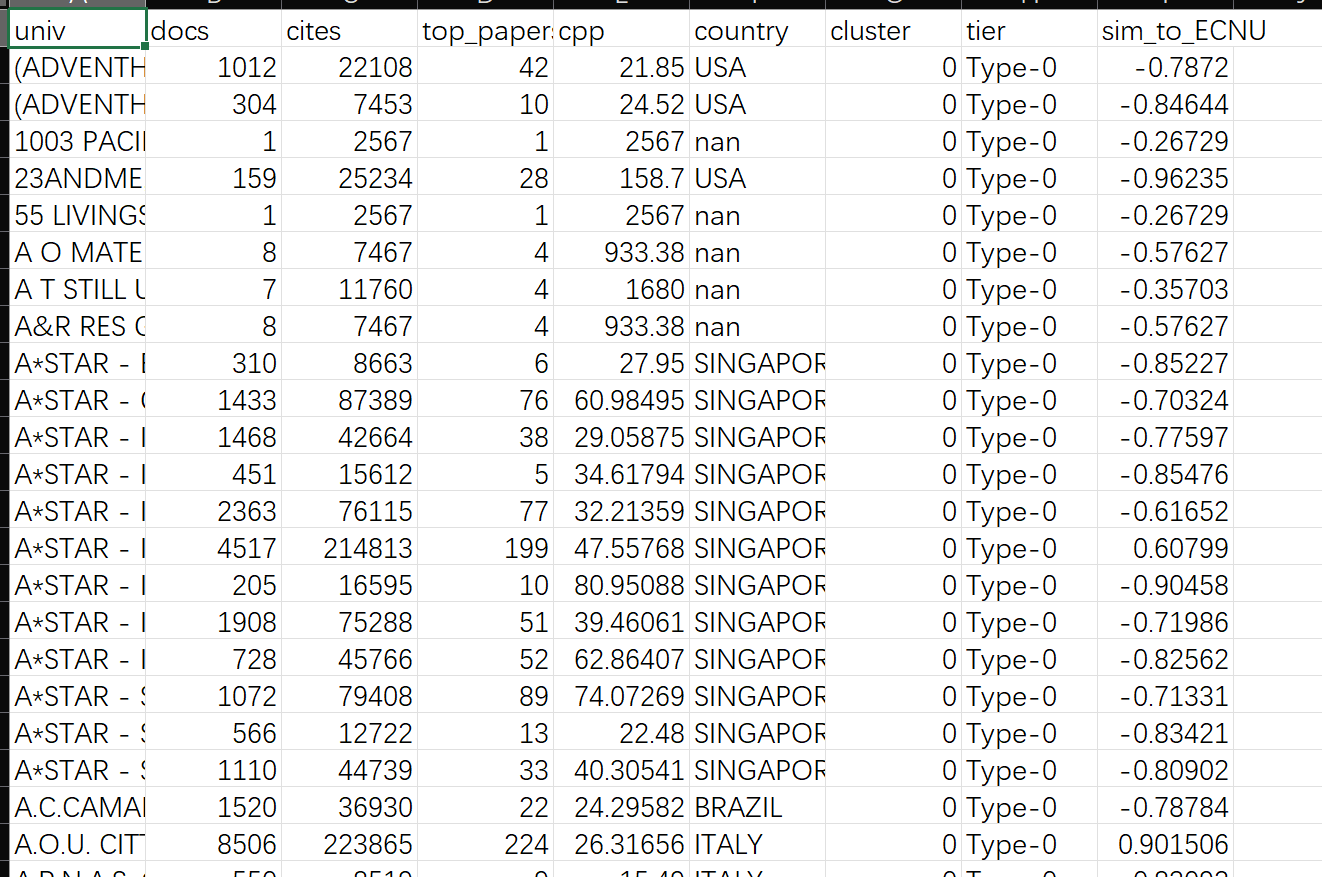
合并后共生成 outputs/esi\_all.csv，包含全球多学科共 34,000+ 条数据。

## 三、问题 8：全球高校聚类分析

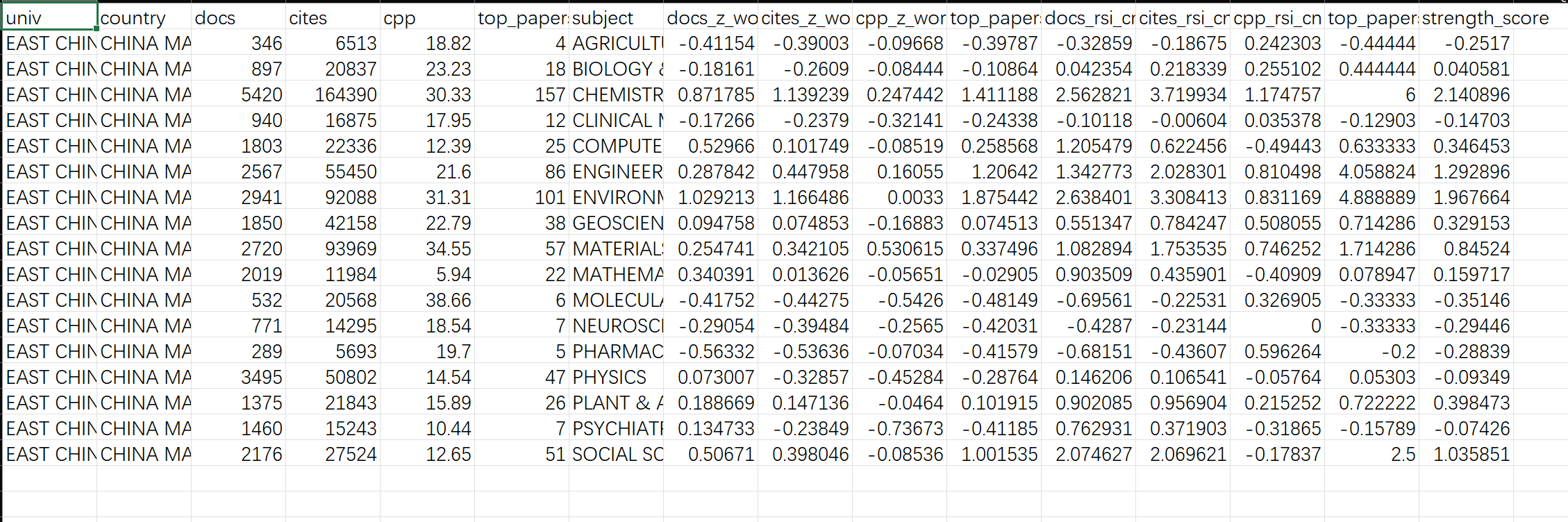
根据各高校在不同学科的论文数、被引数、平均被引及高被引论文比例，使用 KMeans 聚类方法对高校进行聚类。

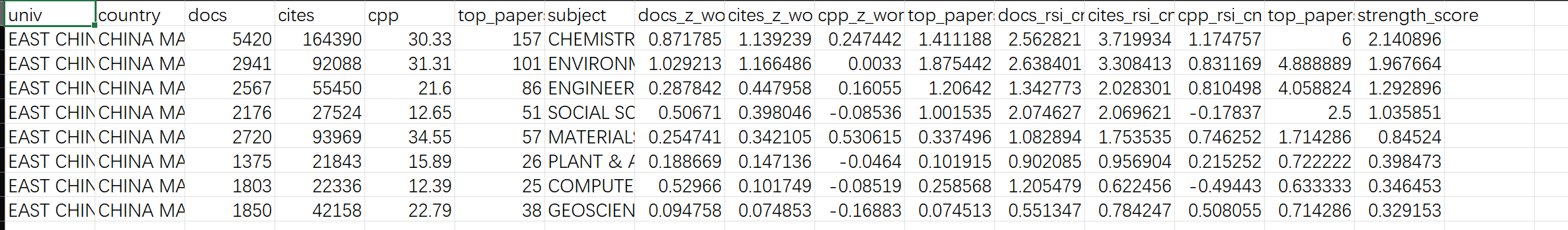
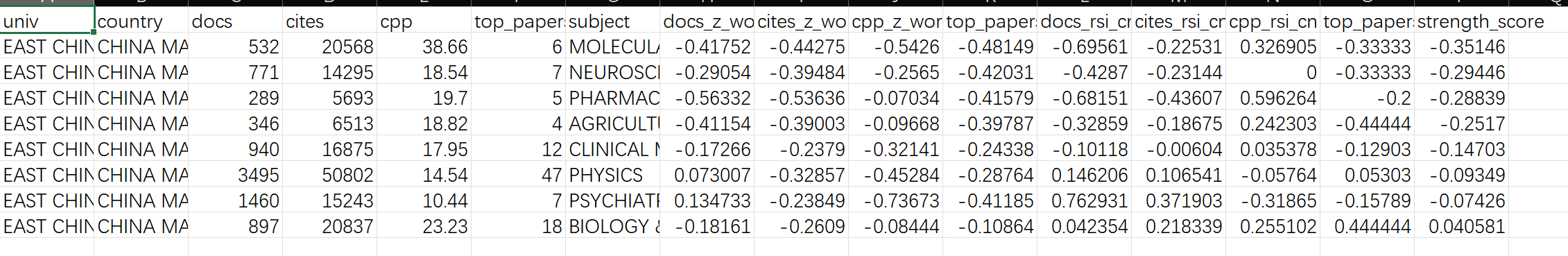
系统自动选择最佳聚类数 k=4（轮廓系数 0.856），说明高校总体可划分为四类：  
① 顶尖科研型大学（Type-0）：论文数与被引数均居全球前列，如哈佛大学、中国科学院；  
② 高影响研究型大学（Type-1）：高被引比例高，代表如斯坦福大学、牛津大学；  
③ 区域研究型大学（Type-2）：学科数量较全但总体引用较低，如东南大学、复旦大学；  
④ 专业性高校（Type-3）：集中于少数学科，代表如农业、地学类院校。



通过计算 ECNU 与各校的余弦相似度发现，华东师范大学最接近 Type-2 类（区域研究型高校），与中山大学、华中师范大学的学科分布相似。

## 四、问题 9：华东师范大学学科画像

通过 ecnu\_profile\_by\_subject.csv、ecnu\_strength\_subjects.csv、ecnu\_weak\_subjects.csv，从多个维度构建华东师范大学的学科画像。

优势学科（Top Strengths）：数学、地球科学、教育学、环境/生态学，在全球同类学科中处于领先地位。  
弱势学科（Weak Subjects）：工程学、计算机科学、临床医学，引用率与高被引论文比例较低。

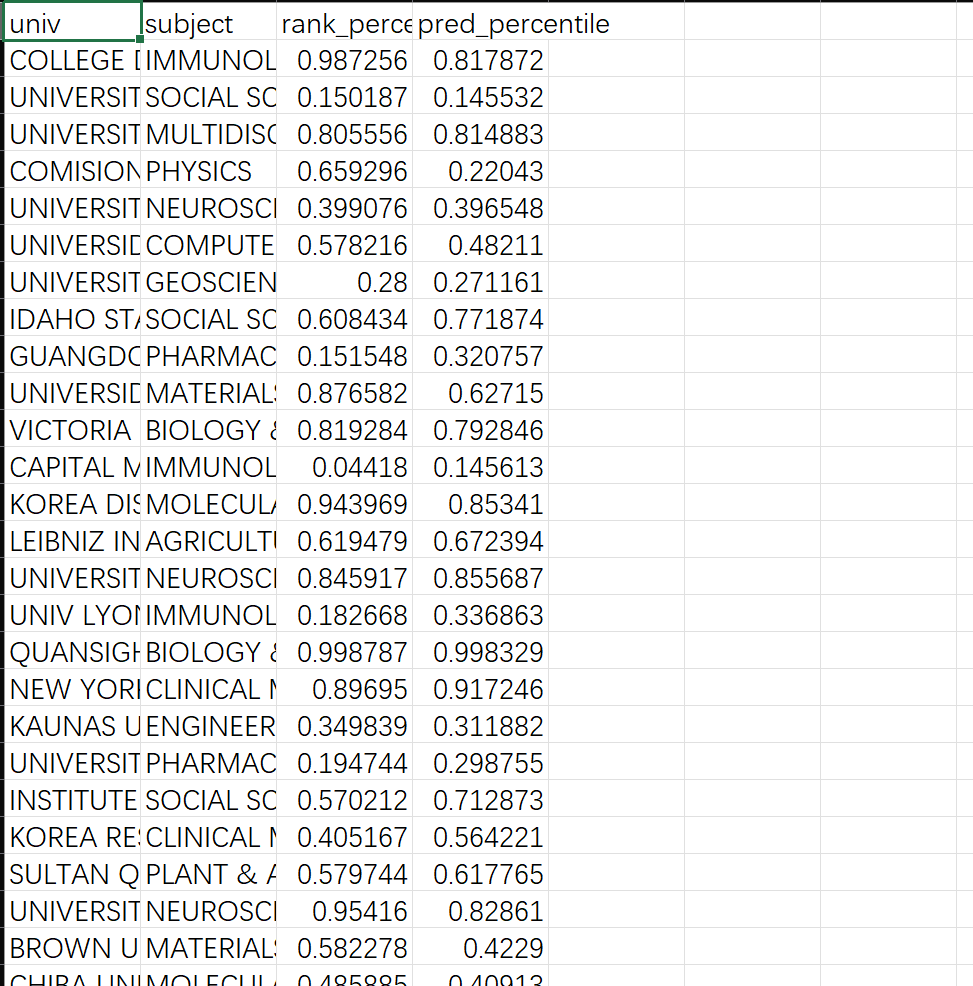
各学科的综合得分通过世界均值标准化（Z-Score）与中国中位数对比（RSI）计算，体现 ECNU 在基础科学领域具有稳定优势，在应用科学领域仍有提升空间。

## 五、问题 10：学科排名预测模型

基于 RandomForestRegressor 随机森林模型构建学科排名预测系统。

特征选取：log(论文数)、log(被引数)、平均被引（CPP）、高被引论文数、Top论文比。  
样本划分：60%训练、20%验证、20%测试。  
评价指标：MAE、RMSE、Spearman 相关系数。

结果：  
验证集 MAE=0.045 RMSE=0.066 Spearman=0.842  
测试集 MAE=0.046 RMSE=0.068 Spearman=0.837  
说明模型能较好预测高校在各学科的相对位置，预测稳定性高。

特征重要性排序：log\_cites > cpp > top\_papers > log\_docs > top\_ratio，表明被引次数与平均被引水平是影响学科排名的关键因素。

## 六、结论与思考

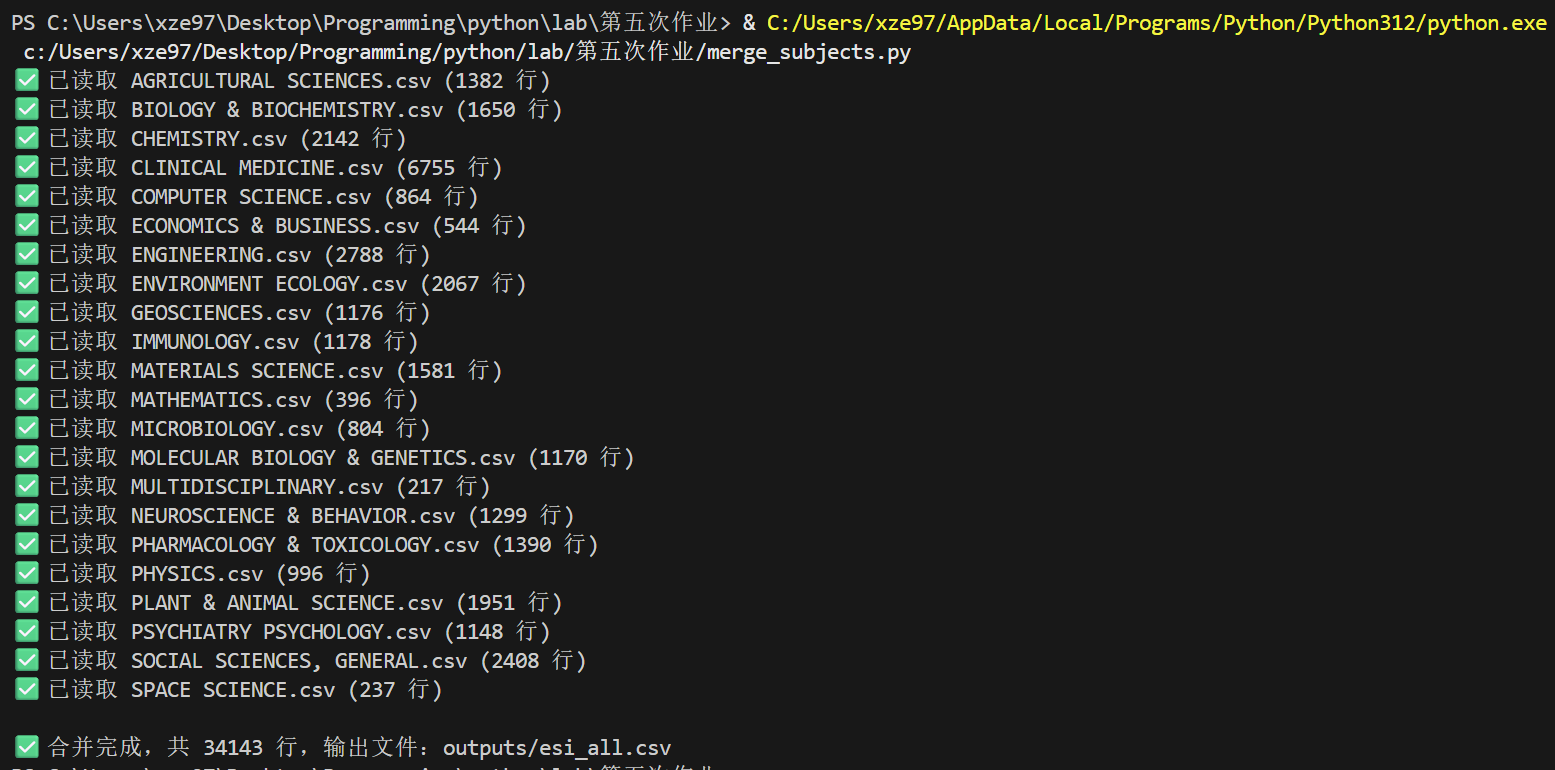
本实验通过数据清洗、聚类分析、学校画像与机器学习预测四个步骤，系统性地揭示了全球高校学科发展的分布规律与特征。

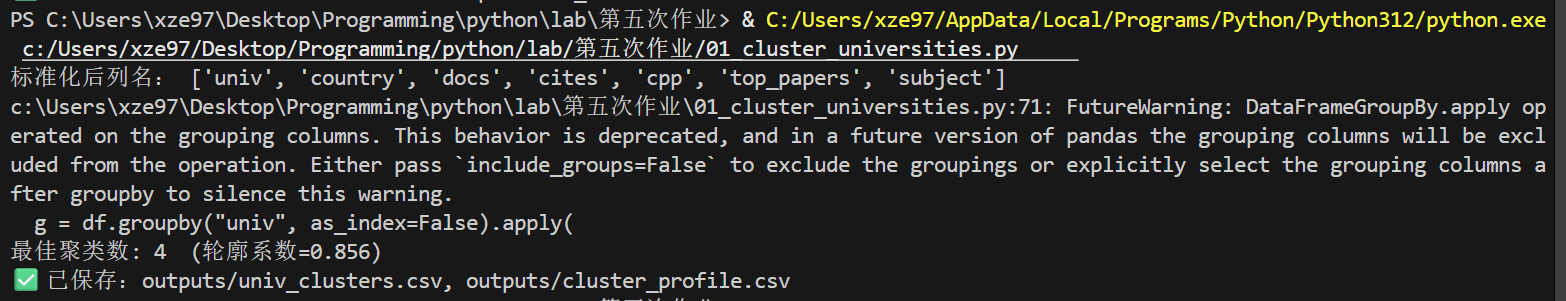
华东师范大学整体处于区域研究型高校水平，在数学、地学、教育等基础学科表现突出，但在工程与医学等应用学科仍存在提升空间。

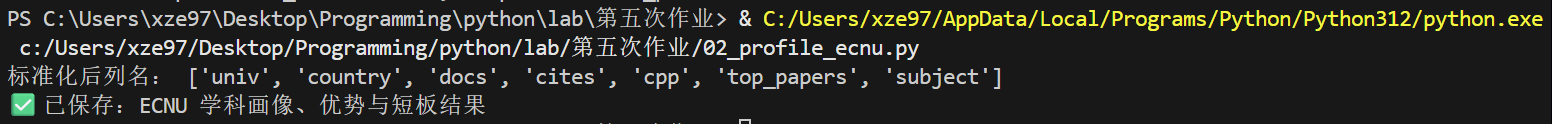
未来可结合时间序列数据进行趋势预测，或使用深度学习模型进一步提高预测精度。

## 七、实验截图

merge\_subjects.py运行截图



01\_cluster\_universities.py运行截图

02\_profile\_ecnu.py运行截图

03\_rank\_model.py运行截图