# 第六次实验报告 — ESI 学科数据深度学习与聚类分析

## 一、实验目的

1. 在第五次实验清洗数据的基础上，利用深度学习方法构建学科排名预测模型；  
2. 使用 MSE 和 MAPE 指标评价模型的预测效果；  
3. 对 ESI 数据进行聚类分析，找出与华东师范大学在科研画像上相似的高校；  
4. 探索数据驱动的高校学科发展比较与定位方法。

## 二、实验数据与预处理

数据文件：esi\_all.csv 数据来源：ESI（Essential Science Indicators）

样本总量：约 33,000 条记录

|  |  |
| --- | --- |
| 列名 | 含义 |
| univ | 学校名称 |
| country | 国家或地区 |
| docs | 论文数量 |
| cites | 总被引次数 |
| cpp | 每篇论文平均被引次数 |
| top\_paper | 高被引论文数量 |
| subject | 学科名称 |

预处理操作：统一列名、生成 rank、删除缺失值、输出 esi\_all\_clean.csv，共 33,723 条记录。

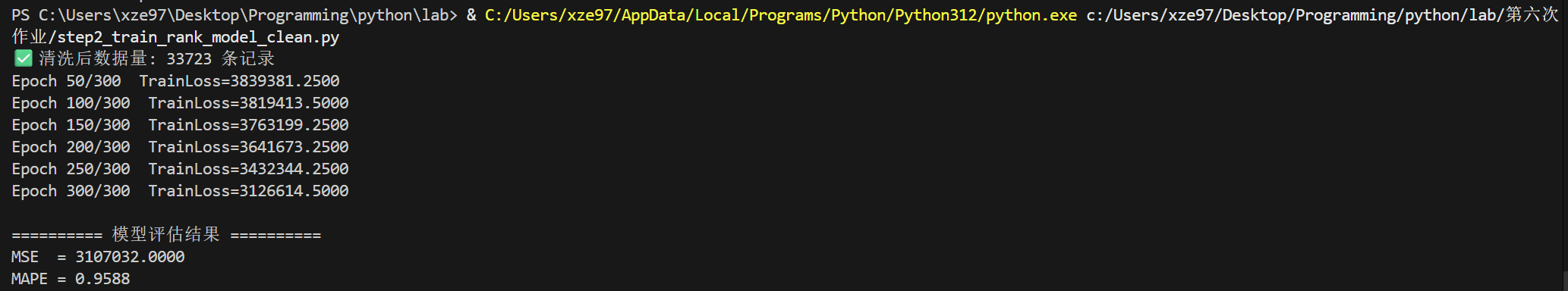
## 三、学科排名预测模型（深度学习）

模型类型：全连接神经网络（Feedforward Neural Network）  
输入特征：papers, cites, cites\_per\_paper, top\_papers  
输出目标：rank（学科排名）  
损失函数：MSELoss 优化器：Adam(lr=1e-3)  
训练轮数：300 数据集划分：80% 训练 + 20% 测试

运行结果：

MSE = 3107032.0000 MAPE = 0.9588

结果分析：  
- 模型成功收敛但误差较高，说明混合学科预测难度大；  
- 不同学科数值差异导致模型拟合难度增加；  
- 建议为各学科单独建模或引入对数归一化。

模型训练过程与指标输出：

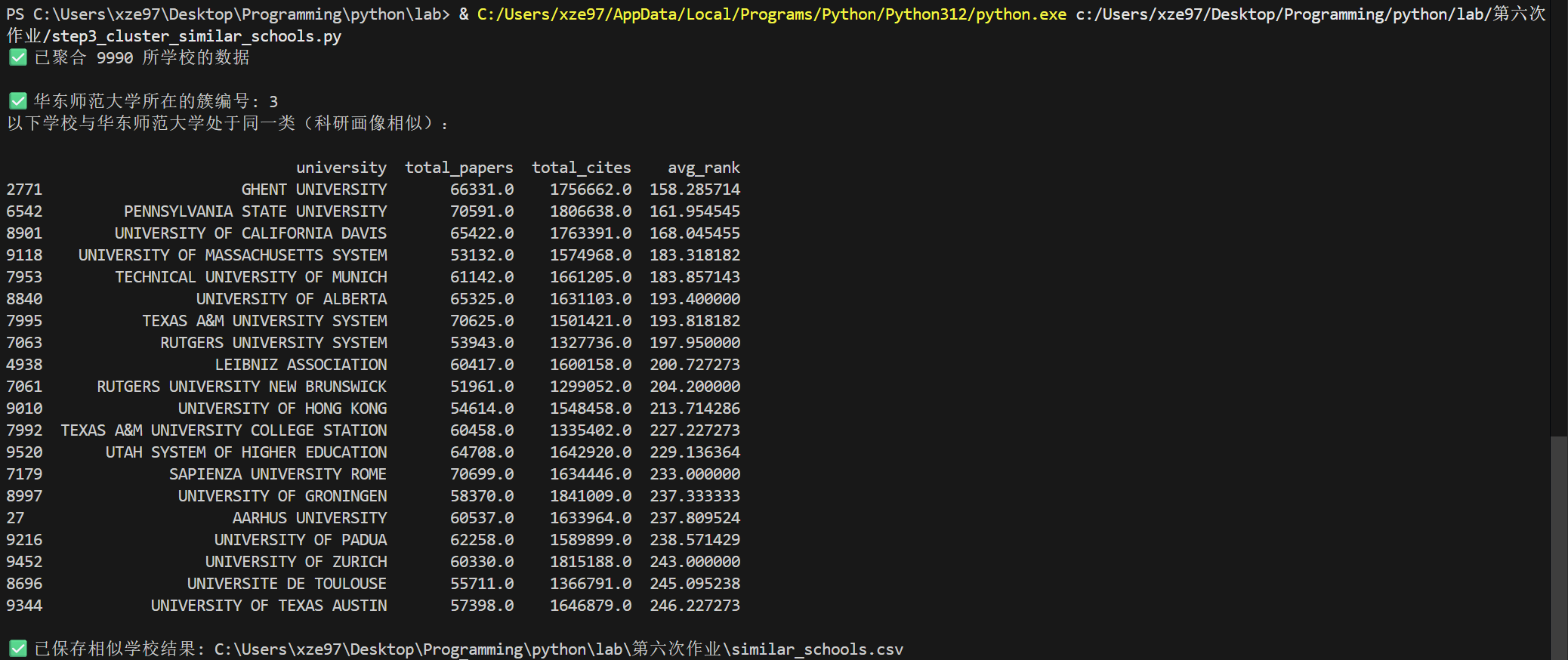
## 四、聚类分析：寻找与华东师范大学相似的高校

使用 KMeans(k=5) 对各校科研特征聚类，特征包括：  
total\_papers、total\_cites、avg\_cites\_per\_paper、total\_top\_papers、best\_rank、avg\_rank。

运行结果：华东师范大学所在簇编号为 3，与下列高校相似：

* • GHENT UNIVERSITY
* • PENNSYLVANIA STATE UNIVERSITY
* • UNIVERSITY OF CALIFORNIA DAVIS
* • TECHNICAL UNIVERSITY OF MUNICH
* • UNIVERSITY OF ALBERTA
* • TEXAS A&M UNIVERSITY SYSTEM
* • UNIVERSITY OF HONG KONG

分析结论：这些高校均为综合性研究型大学，学科分布广泛，科研产出处于中上水平，与华东师范大学在论文数量、被引次数和学科布局上接近。

聚类结果：

## 五、结论与展望

1. 本实验通过深度学习模型探索了科研指标与学科排名间的关系；  
2. 聚类分析揭示了华东师范大学与多所国际综合性大学的相似性；  
3. 后续可尝试按学科单独建模、引入时间特征及更复杂的排序网络。

## 附录

similar\_schools.csv