实验五：探索性分析与数据建模

8. 请结合这些学科排名数据，分析全球高校可以大致分为哪几类?并且分析出与华师大类似的高校?

数据是每个学科一个表格，适合观察某个学科哪些高校厉害

但是如果想将全球高校分类，应该观察**同一个高校的不同学科**，数据的组织形式是不适合这种观察的

我的思路是把所有表格合并成一个二维表格，每一行表示一个高校，每一列表示一个学科

某行某列的元素代表**某个高校在某个学科的排名**

手动构建这个二维表格的工作量是巨大的，我们肯定要写一个python程序

但是原数据的表格第一行和最后一行应该先清理掉，以便python程序处理

这一步很重要，第8题和第10题都会用到

先打开一个csv看一下，发现第1行是说明文字，第2行才是表头：

	A	B	C	D	E	F	G	H	I	J	K	L
1	Indicators	Results List:	Institutions	Filter Results By:		ResearchFields	Filter Value(s):	AGRICULTURAL SCIENCES			Show:	Top
2		Institution	Countries/	Web of Sc	Cites	Cites/Pape	Top Papers					
3	1	CHINESE /	CHINA MA	15661	332254	21.22	251					
4	2	CHINESE /	CHINA MA	12222	223855	18.32	198					
5	3	UNITED S	USA	12564	220644	17.56	105					
6	4	CHINA AG	CHINA MA	10052	207779	20.67	166					
7	5	INRAE	FRANCE	9314	187838	20.17	118					
8	6	JIANGNAI	CHINA MA	8027	181325	22.59	166					
9	7	NORTHW	CHINA MA	9024	180634	20.02	162					
10	8	WAGENIN	NETHERLA	6819	172734	25.33	152					
11	9	CONSEJO	SPAIN	7656	167593	21.89	94					
12	10	NANJING	CHINA MA	6480	137921	21.28	99					
13	11	UNIVERSI	USA	6175	129479	20.97	110					
14	12	UNIVERSI	CHINA MA	5950	127682	21.46	95					
15	13	ZHEJIANG	CHINA MA	5313	125715	23.66	148					
16	14	EGYPTIAN	EGYPT	7861	117405	14.94	103					
17	15	INDIAN C	INDIA	12753	111262	8.72	66					
18	16	SOUTH CI	CHINA MA	3524	108282	30.73	117					
19	17	CGIAR	FRANCE	5262	107193	20.37	69					
20	18	HUAZHOI	CHINA MA	4855	97482	20.08	68					
21	19	UNIVERSI	BRAZIL	6169	95242	15.44	35					
22	20	JIANGSU U	CHINA MA	4084	89040	21.8	176					
23	21	STATE UN	USA	5172	86527	16.73	62					
24	22	CENTRE N	FRANCE	4113	81348	19.78	60					
25	23	CORNELL	USA	3551	78744	22.18	59					

最后一行是 Copyright ?2025 Clarivate，也应该去掉：

	A	B	C	D	E	F	G	H	I	J	K	L
1359	1356	MAE FAH	THAILAND	236	3627	15.37	2					
1360	1358	NAMIK KE	TURKIYE	352	3625	10.3	1					
1361	1359	GANGNEU	SOUTH K(	268	3621	13.51	1					
1362	1360	TAMPERE	FINLAND	234	3619	15.47	0					
1363	1361	HOSPITAL	CANADA	202	3618	17.91	2					
1364	1362	BRAUNSC	GERMANY	188	3614	19.22	1					
1365	1362	ICAR - INI	INDIA	393	3614	9.2	6					
1366	1364	HUNGARI	HUNGARY	180	3605	20.03	2					
1367	1365	LIAONING	CHINA MA	333	3604	10.82	0					
1368	1366	DAIRYNZ	NEW ZEAI	239	3599	15.06	0					
1369	1367	SMITHSO	USA	209	3598	17.22	0					
1370	1367	BAYER CR	GERMANY	250	3598	14.39	2					
1371	1369	GENERAL	USA	107	3595	33.6	4					
1372	1370	STEFAN C	ROMANIA	161	3591	22.3	5					
1373	1371	KAGOSHI	JAPAN	306	3588	11.73	1					
1374	1372	UNIVERSI	USA	127	3580	28.19	4					
1375	1373	HOPITAL	FRANCE	123	3573	29.05	3					
1376	1373	HEINRICH	GERMANY	136	3573	26.27	4					
1377	1375	UNIVERSI	SOUTH AI	184	3572	19.41	6					
1378	1376	BANGLAD	BANGLAD	170	3567	20.98	2					
1379	1376	NEAR EAS	TURKIYE	204	3567	17.49	2					
1380	1378	GRIFFITH	AUSTRALI	205	3562	17.38	2					
1381	1378	UMM AL-	SAUDI AR	308	3562	11.56	3					
1382	1380	SAINT LO	USA	108	3561	32.97	4					
1383	1381	VIT VELLC	INDIA	144	3558	24.71	2					
1384	Copyright ?2025 Clarivate											
1385												

所以我们先把每个 CSV 的**第一行和最后一行删掉**，再导入

在 Powershell 执行下面命令：

```
$src="D:\ESI\download"; $dst="D:\ESI\clean"; New-Item -ItemType Directory -Force
-Path $dst | Out-Null
Get-ChildItem $src -Filter *.csv | ForEach-Object {
  $lines = Get-Content $_.FullNameGet-ChildItem $src -Filter *.csv
  $lines[1..($lines.Count-2)] | Set-Content -Encoding UTF8 (Join-Path $dst
$_.Name)
}
```

解释一下：

1. src 是原始文件路径，dst 是目标文件路径
2. New-Item 用于创建一个新目录，如果 D:\ESI\clean 不存在，会自动创建
3. Get-ChildItem $src -Filter *.csv 会列出 src 中所有扩展名为 .csv 的文件
4. ForEach-Object 遍历每个文件
5. $lines = Get-Content $_.FullName 会把整个文件的内容读成一个字符串数组，每一行是一个元素
6. $lines[1..($lines.Count-2)] 是数组切片语法，表示从第二行到倒数第二行，删除文件的第一行和最后一行
7. Set-Content 把上一步得到的“去掉第一行的内容”写入到新的文件中

可以看到，第一行已经成功去掉了：

	A	B	C	D	E	F	G	H	I	J	K	L
1		Institution	Countries/	Web of Sc	Cites	Cites/Pape	Top Papers					
2	1	CHINESE /	CHINA MA	15661	332254	21.22	251					
3	2	CHINESE /	CHINA MA	12222	223855	18.32	198					
4	3	UNITED S	USA	12564	220644	17.56	105					
5	4	CHINA AC	CHINA MA	10052	207779	20.67	166					
6	5	INRAE	FRANCE	9314	187838	20.17	118					
7	6	JIANGNAI	CHINA MA	8027	181325	22.59	166					
8	7	NORTHW	CHINA MA	9024	180634	20.02	162					
9	8	WAGENIN	NETHERLA	6819	172734	25.33	152					
10	9	CONSEJO	SPAIN	7656	167593	21.89	94					
11	10	NANJING	CHINA MA	6480	137921	21.28	99					
12	11	UNIVERSI	USA	6175	129479	20.97	110					
13	12	UNIVERSI	CHINA MA	5950	127682	21.46	95					
14	13	ZHEJIANG	CHINA MA	5313	125715	23.66	148					
15	14	EGYPTIAN	EGYPT	7861	117405	14.94	103					
16	15	INDIAN C	INDIA	12753	111262	8.72	66					
17	16	SOUTH CI	CHINA MA	3524	108282	30.73	117					
18	17	CGIAR	FRANCE	5262	107193	20.37	69					
19	18	HUAZHOI	CHINA MA	4855	97482	20.08	68					
20	19	UNIVERSII	BRAZIL	6169	95242	15.44	35					
21	20	JIANGSU U	CHINA MA	4084	89040	21.8	176					
22	21	STATE UN	USA	5172	86527	16.73	62					
23	22	CENTRE N	FRANCE	4113	81348	19.78	60					
24	23	CORNELL	USA	3551	78744	22.18	59					
25	24	UNIVERSI	USA	3908	78346	20.05	64					

最后一行也成功去掉了：

	A	B	C	D	E	F	G	H	I	J	K	L
1366	1365	LIAONING	CHINA MA	333	3604	10.82	0					
1367	1366	DAIRYNZ	NEW ZEAI	239	3599	15.06	0					
1368	1367	SMITHSOI	USA	209	3598	17.22	0					
1369	1367	BAYER CR	GERMANY	250	3598	14.39	2					
1370	1369	GENERAL	USA	107	3595	33.6	4					
1371	1370	STEFAN C	ROMANIA	161	3591	22.3	5					
1372	1371	KAGOSHI	JAPAN	306	3588	11.73	1					
1373	1372	UNIVERSI	USA	127	3580	28.19	4					
1374	1373	HOPITAL I	FRANCE	123	3573	29.05	3					
1375	1373	HEINRICH	GERMANY	136	3573	26.27	4					
1376	1375	UNIVERSI	SOUTH AF	184	3572	19.41	6					
1377	1376	BANGLAD	BANGLAD	170	3567	20.98	2					
1378	1376	NEAR EAS	TURKIYE	204	3567	17.49	2					
1379	1378	GRIFFITH	AUSTRALI	205	3562	17.38	2					
1380	1378	UMM AL-	SAUDI AR	308	3562	11.56	3					
1381	1380	SAINT LO	USA	108	3561	32.97	4					
1382	1381	VIT VELLC	INDIA	144	3558	24.71	2					
1383												
1384												
1385												
1386												
1387												
1388												
1389												
1390												
1391												

下面来构建二维表格，重复一下我的思路：

> 把所有表格合并成一个二维表格，每一行表示一个高校，每一列表示一个学科
>
> 某行某列的元素代表**某个高校在某个学科的排名**

我们读取clean目录下的所有CSV文件，把二维表格输出到一个CSV文件中：

```
INPUT_DIR = (Path(__file__).resolve().parent / "clean")
OUTPUT_CSV = (Path(__file__).resolve().parent / "8_build.csv")
```

先写一个函数，列出所有CSV文件：

```
def list_subject_files(root: Path) -> list[Path]:
    files = sorted(root.glob("*.csv"))
    if not files:
        raise FileNotFoundError(f"No CSV files found in {root}")
    return files
```

对于每个学科表，我们只关心大学名称、学科名称、学科排名这三个信息

学科排名：rank 取文件的第一列（清洗后的名次列）

大学名称：university 取 'Institutions'

学科名称：subject 用文件名（去掉扩展名）

对于每个学科表，处理后返回一个 pd.DataFrame，包含三列：['university', 'subject', 'rank']

读入 CSV，找到表示名次的第一列，指定表示学校名称的 Institutions 列：

```
df = pd.read_csv(csv_path, dtype=str)
rank_col = df.columns[0]
inst_col = "Institutions"

out = pd.DataFrame({
    "university": df[inst_col].astype(str).str.strip(),
    "rank": pd.to_numeric(df[rank_col], errors="coerce"),
}).dropna(subset=["university", "rank"])

out["rank"] = out["rank"].astype(int)
out["subject"] = csv_path.stem.strip()
return out[["university", "subject", "rank"]]
```

每个学科都会返回一个 pd.DataFrame，将所有 pd.DataFrame 合并：

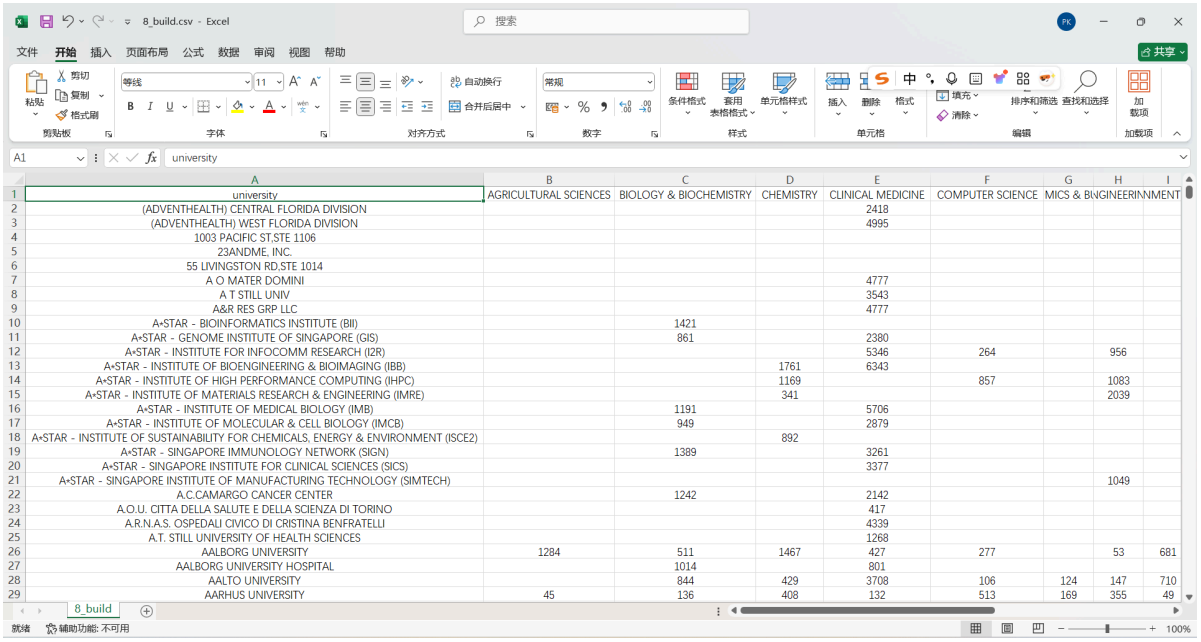
```
frames = [read_subject(p) for p in list_subject_files(input_dir)]
long_df = pd.concat(frames, ignore_index=True)
```

最后，按行、列排序：

```
return pivot.sort_index().sort_index(axis=1)
```

以上预处理的完整代码请老师阅读 8_build.py

运行后，得到二维表格，取名叫 8_build.csv，同样在附件中：



	university	AGRICULTURAL SCIENCES	BIOLOGY & BIOCHEMISTRY	CHEMISTRY	CLINICAL MEDICINE	COMPUTER SCIENCE	MICS & B	NGINEERI	NMENT
2	(ADVENTHEALTH) CENTRAL FLORIDA DIVISION				2418				
3	(ADVENTHEALTH) WEST FLORIDA DIVISION				4995				
4	1003 PACIFIC ST,STE 1106								
5	23ANDME, INC.								
6	55 LIVINGSTON RD,STE 1014								
7	A O MATER DOMINI				4777				
8	A T STILL UNIV				3543				
9	A&R RES GRP LLC				4777				
10	A*STAR - BIOINFORMATICS INSTITUTE (BII)		1421						
11	A*STAR - GENOME INSTITUTE OF SINGAPORE (GIS)		861		2380				
12	A*STAR - INSTITUTE FOR INFOCOMM RESEARCH (I2R)				5346	264		956	
13	A*STAR - INSTITUTE OF BIOENGINEERING & BIOIMAGING (IBB)			1761	6343				
14	A*STAR - INSTITUTE OF HIGH PERFORMANCE COMPUTING (IHPC)			1169		857		1083	
15	A*STAR - INSTITUTE OF MATERIALS RESEARCH & ENGINEERING (IMRE)			341				2039	
16	A*STAR - INSTITUTE OF MEDICAL BIOLOGY (IMB)		1191		5706				
17	A*STAR - INSTITUTE OF MOLECULAR & CELL BIOLOGY (IMCB)		949		2879				
18	A*STAR - INSTITUTE OF SUSTAINABILITY FOR CHEMICALS, ENERGY & ENVIRONMENT (ISCE2)			892					
19	A*STAR - SINGAPORE IMMUNOLOGY NETWORK (SIGN)		1389		3261				
20	A*STAR - SINGAPORE INSTITUTE FOR CLINICAL SCIENCES (SICS)				3377				
21	A*STAR - SINGAPORE INSTITUTE OF MANUFACTURING TECHNOLOGY (SIMTECH)							1049	
22	A.C.CAMARGO CANCER CENTER		1242		2142				
23	A.O.U. CITTA DELLA SALUTE E DELLA SCIENZA DI TORINO				417				
24	A.R.N.A.S. OSPEDALI CIVICO DI CRISTINA BENFRATELLI				4339				
25	A.T. STILL UNIVERSITY OF HEALTH SCIENCES				1268				
26	AALBORG UNIVERSITY	1284	511	1467	427	277		53	681
27	AALBORG UNIVERSITY HOSPITAL		1014		801				
28	AALTO UNIVERSITY		844	429	3708	106	124	147	710
29	AARHUS UNIVERSITY	45	136	408	132	513	169	355	49

空白元素是正常的，代表这个高校在这个学科排名中**没有上榜**

有了这个表格，我们就可以直观地将全球高校进行分类

全球高校可以大致分为哪几类?

综合类、理工类、医学类、文科-社科类

这还不够，为了严格地将所有高校分类，我需要通过观察表格数据，设计一个**判定标准**

综合类：至少两大组各有 ≥2 门学科进入 Top200，且总覆盖 ≥10 门

理工类：理工组 Top100 学科数 ≥2，且明显多于其他两组（约 ≥2 倍），生命/社科 Top100 极少（≤1）

医学类：生命-医学组 Top100 学科数 ≥2，且明显多于其他两组（约 ≥2 倍）

文科-社科类：社科组有 Top200 学科 ≥1，且理工/生命医学 Top100 皆为 0

未分类：其余的

我将每个类别的所有高校整理成了表格，在附件中供老师查看：

universities_engineering_science.csv

universities_comprehensive.csv

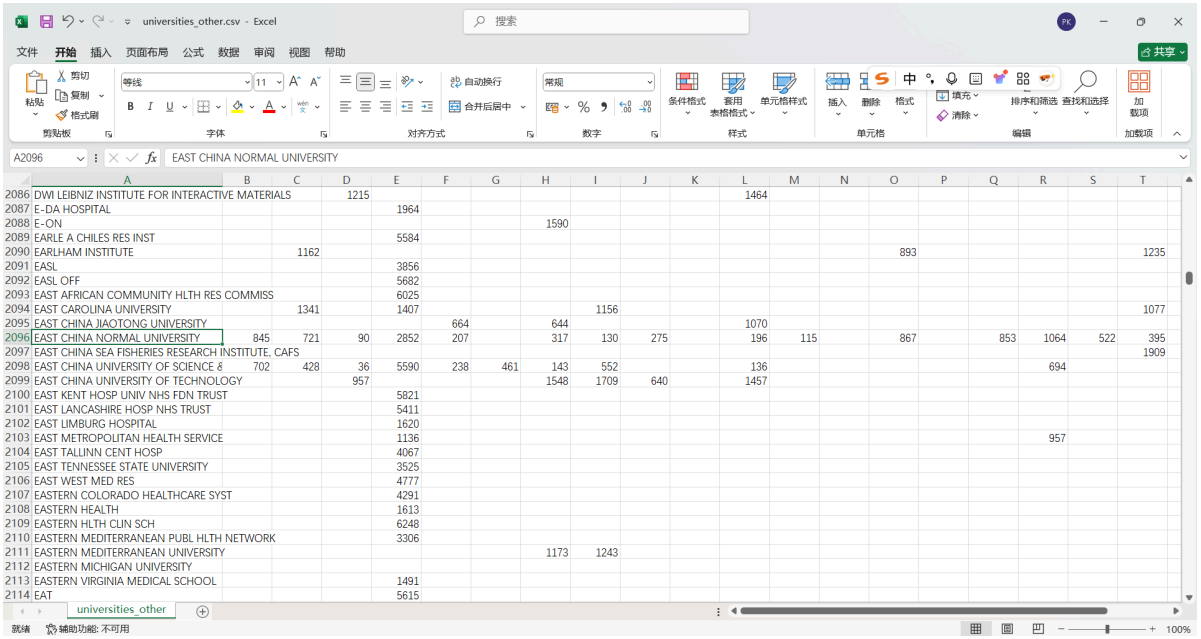
universities_other.csv

universities_social.csv

universities_medical.csv

哪些高校与华师大类似?

按照刚才的分类标准，华师大属于“未分类”的高校，所以无法直接找到同类



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
2086	DWI LEIBNIZ INSTITUTE FOR INTERACTIVE MATERIALS			1215								1464								
2087	E-DA HOSPITAL				1964															
2088	E-ON							1590												
2089	EARLE A CHILES RES INST				5584															
2090	EARLHAM INSTITUTE		1162												893					1235
2091	EASL				3856															
2092	EASL OFF				5682															
2093	EAST AFRICAN COMMUNITY HLTH RES COMMISS				6025															
2094	EAST CAROLINA UNIVERSITY		1341		1407				1156											1077
2095	EAST CHINA JIAOTONG UNIVERSITY					664		644				1070								
2096	EAST CHINA NORMAL UNIVERSITY	845	721	90	2852	207		317	130	275		196	115		867		853	1064	522	395
2097	EAST CHINA SEA FISHERIES RESEARCH INSTITUTE, CAFS																			1909
2098	EAST CHINA UNIVERSITY OF SCIENCE &	702	428	36	5590	238	461	143	552			136						694		
2099	EAST CHINA UNIVERSITY OF TECHNOLOGY			957				1548	1709	640		1457								
2100	EAST KENT HOSP UNIV NHS FDN TRUST				5821															
2101	EAST LANCASHIRE HOSP NHS TRUST				5411															
2102	EAST LIMBURG HOSPITAL				1620															
2103	EAST METROPOLITAN HEALTH SERVICE				1136													957		
2104	EAST TALLINN CENT HOSP				4067															
2105	EAST TENNESSEE STATE UNIVERSITY				3525															
2106	EAST WEST MED RES				4777															
2107	EASTERN COLORADO HEALTHCARE SYST				4291															
2108	EASTERN HEALTH				1613															
2109	EASTERN HLTH CLIN SCH				6248															
2110	EASTERN MEDITERRANEAN PUBL HLTH NETWORK				3306															
2111	EASTERN MEDITERRANEAN UNIVERSITY							1173	1243											
2112	EASTERN MICHIGAN UNIVERSITY																			
2113	EASTERN VIRGINIA MEDICAL SCHOOL				1491															
2114	EAT				5615															

现在的问题是，如果给定两个大学的各学科排名向量，如何计算相似度？

首先，排名应该转化为**得分**，因为名次越小越强，而余弦相似度默认“数值越大越强”：

得分介于0到1，排名越大，得分越小

$$\text{score} = 1 - \frac{rank - 1}{\max\_rank\_s - 1} \quad \in [0, 1]$$

未上榜怎么处理？不是 0 分，而是设成“比榜尾还低一点”的极低分：

$$\text{score\_for\_NaN} = \frac{1}{\max\_rank\_s + 1}$$

这样，处理后的两个高校的向量就可以计算相似度了，只需要计算**余弦相似度**

首先，我们需要从 8_build.csv 文件中加载数据，并且将名次列从字符串转换为数值类型：

```
df = pd.read_csv(CSV)
u, subs = df.columns[0], df.columns[1:]
df[subs] = df[subs].apply(pd.to_numeric, errors="coerce")
```

接下来，我们将每个学科的排名转换为得分：

```
m = df[subs].max(skipna=True)              # 各学科最大名次 max_rank_s
den = (m - 1).replace(0, np.nan)           # 防止除零
score = 1 - (df[subs] - 1).div(den, axis=1) # 式(1)
score = score.clip(0, 1).fillna(1 / (m + 1)) # 未上榜用极低分
S = pd.concat([df[[u]], score], axis=1)    # 得分矩阵（行=高校，列=学科）
```

现在我们需要从得分矩阵中提取出 East China Normal University 的得分向量：

```
mask = S[u].str.contains(TARGET, case=False, na=False) | S[u].str.contains("华东师范", na=False)
if not mask.any():
    raise SystemExit(f"target not found: {TARGET}")
t = S.loc[mask, subs].iloc[0].to_numpy(float)
```

我们使用余弦相似度来衡量两所高校在各个学科得分上的相似程度：

```
M = S.loc[~mask, subs].to_numpy(float)  # 取其他高校的得分矩阵
names = S.loc[~mask, u].to_numpy()      # 取其他高校名称

denom = np.linalg.norm(M, axis=1) * np.linalg.norm(t)  # 计算分母：向量的模
sims = (M @ t) / denom  # 计算余弦相似度
ok = np.isfinite(sims)  # 处理无效值
names, sims = names[ok], sims[ok]  # 过滤无效的相似度
```

最后，我们将计算出的相似度按照降序排序，并输出到文件：

```
order = np.argsort(-sims)  # 按相似度从高到低排序
with open(OUT, "w", encoding="utf-8") as f:
    f.write(f"Similar to EAST CHINA NORMAL UNIVERSITY (N={len(order)})\n")
    for i, idx in enumerate(order, 1):
        f.write(f"{i:4d}. {names[idx]} | sim={sims[idx]:.4f}\n")
```

完整代码在附件中，叫 8_ecnu_find.py

运行代码后，所有高校按照与华师大的**相似度从高到低**，都打印出来了，完整结果见 8_ecnu_find.txt

可见，滑铁卢大学、北京师范大学、电子科技大学与华师大最相似

Similar to EAST CHINA NORMAL UNIVERSITY (N=9989)
1. UNIVERSITY OF WATERLOO | sim=0.9566
2. BEIJING NORMAL UNIVERSITY | sim=0.9513
3. UNIVERSITY OF ELECTRONIC SCIENCE & TECHNOLOGY OF CHINA | sim=0.9494
4. CHONGQING UNIVERSITY | sim=0.9463
5. UNIVERSITY OF VICTORIA | sim=0.9422
6. NORWEGIAN UNIVERSITY OF SCIENCE & TECHNOLOGY (NTNU) | sim=0.9414
7. NANJING NORMAL UNIVERSITY | sim=0.9404
8. SHENZHEN UNIVERSITY | sim=0.9356
9. ISLAMIC AZAD UNIVERSITY | sim=0.9345
10. SOUTHWEST UNIVERSITY - CHINA | sim=0.9338
11. UNIVERSITY OF WOLLONGONG | sim=0.9333
12. UNIVERSITY OF BASQUE COUNTRY | sim=0.9323
13. GEORGIA INSTITUTE OF TECHNOLOGY | sim=0.9296
14. UNIVERSITY OF DELAWARE | sim=0.9290
15. UNIVERSITY OF QUEBEC | sim=0.9278
16. VIRGINIA POLYTECHNIC INSTITUTE & STATE UNIVERSITY | sim=0.9273
17. DALIAN UNIVERSITY OF TECHNOLOGY | sim=0.9273
18. UNIVERSITY OF GRANADA | sim=0.9261
19. KING ABDULAZIZ UNIVERSITY | sim=0.9233
20. LANZHOU UNIVERSITY | sim=0.9221
21. SOUTH CHINA NORMAL UNIVERSITY | sim=0.9211
22. CITY UNIVERSITY OF HONG KONG | sim=0.9209
23. NORTH CAROLINA STATE UNIVERSITY | sim=0.9186
24. PURDUE UNIVERSITY | sim=0.9171
25. PURDUE UNIVERSITY SYSTEM | sim=0.9159
26. UNIVERSITE DE RENNES | sim=0.9158
27. DELFT UNIVERSITY OF TECHNOLOGY | sim=0.9151
28. UNIVERSITY OF CALIFORNIA RIVERSIDE | sim=0.9143
29. MINIST EDUC | sim=0.9139
30. TECHNICAL UNIVERSITY OF BERLIN | sim=0.9119
31. PRES EUROPEAN UNIV BRETAGNE | sim=0.9113
32. UNIVERSITE DE LORRAINE | sim=0.9113
33. NANJING UNIVERSITY | sim=0.9100
34. COMMUNAUTE UNIVERSITE GRENOBLE ALPES | sim=0.9094
35. UNIVERSITY OF SCIENCE & TECHNOLOGY OF CHINA, CAS | sim=0.9090
36. POLISH ACADEMY OF SCIENCES | sim=0.9089
37. UNIVERSITE GRENOBLE ALPES (UGA) | sim=0.9088

9. 请通过探索性分析的方式，对华东师范大学做一个学科画像?用尽可能多的角度去做。

承接上次作业的第五题：通过写SQL语句，获取华东师范大学在各个学科中的排名

第五题的结果文件（命名为 9_ecnu_data），刚好可以作为本题的数据来源：

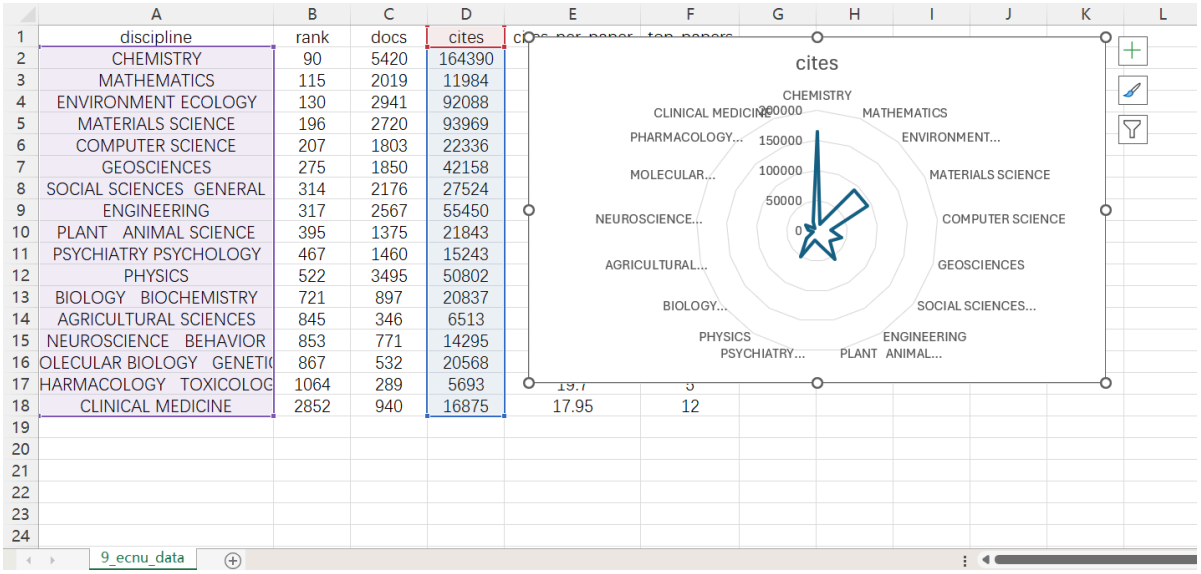
	A	B	C	D	E	F	G
1	discipline	rank	docs	cites	cites_per_paper	top_papers	
2	CHEMISTRY	90	5420	164390	30.33	157	
3	MATHEMATICS	115	2019	11984	5.94	22	
4	ENVIRONMENT ECOLOGY	130	2941	92088	31.31	101	
5	MATERIALS SCIENCE	196	2720	93969	34.55	57	
6	COMPUTER SCIENCE	207	1803	22336	12.39	25	
7	GEOSCIENCES	275	1850	42158	22.79	38	
8	SOCIAL SCIENCES GENERAL	314	2176	27524	12.65	51	
9	ENGINEERING	317	2567	55450	21.6	86	
10	PLANT ANIMAL SCIENCE	395	1375	21843	15.89	26	
11	PSYCHIATRY PSYCHOLOGY	467	1460	15243	10.44	7	
12	PHYSICS	522	3495	50802	14.54	47	
13	BIOLOGY BIOCHEMISTRY	721	897	20837	23.23	18	
14	AGRICULTURAL SCIENCES	845	346	6513	18.82	4	
15	NEUROSCIENCE BEHAVIOR	853	771	14295	18.54	7	
16	OLECULAR BIOLOGY GENETI(	867	532	20568	38.66	6	
17	HARMACOLOGY TOXICOLOG	1064	289	5693	19.7	5	
18	CLINICAL MEDICINE	2852	940	16875	17.95	12	
19							

上课的时候，胡老师提到了探索性分析：

如果数据不够直观，我们可以通过**画图**来更直观地看到数据的分布和变化

我选择了柱状图、饼图、雷达图

先聚焦于**被引用次数**，制作一个**雷达图**：

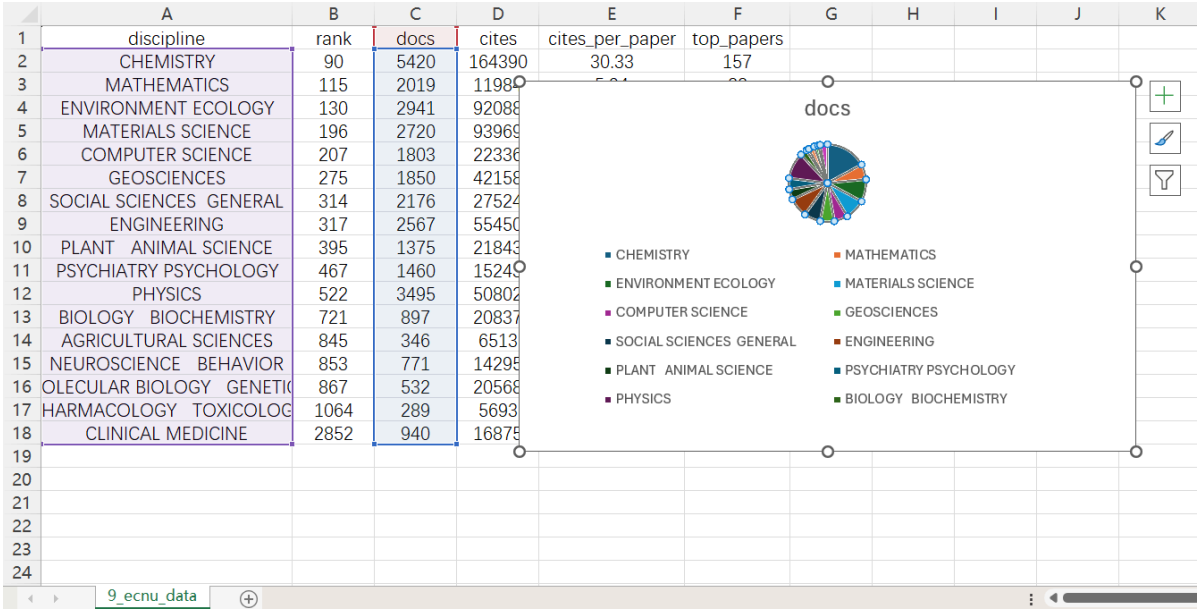


学科画像（1）：

1. 化学学科的被引用次数非常突出，这是华东师范大学最强的学科之一，排名靠前且引用次数显著高于其他学科，说明化学领域的研究成果在国际上有较高的影响力。
2. 临床医学尽管排名较后，但引用次数较高，表明该领域有较强的学术产出和影响力。
3. 分子生物学和药理学，这两个学科的引用次数也较为突出，表明该校在生命科学和医学研究方面有一定的优势。
4. 农业科学虽然排名较高，但是其被引用次数较低，说明该领域的影响力相对较弱。
5. 心理学虽然排名在前列，但是在引用次数上的表现较为一般，表明该学科的学术影响力还有提升空间。
6. **学科多样性**：华东师范大学的学科覆盖范围广泛，涵盖了自然科学（化学、物理、地学等）、生命科学（生物学、医学等）以及社会科学（社会学、心理学等）。通过雷达图可以看出，虽然在一些领域（如生物学、地学、社会学）排名处于中等，但大部分学科的影响力在逐步上升。

7. 排名与引用次数的关系：排名与引用次数之间并不总是成正比，前面的几条可以作为很好的例子
8. 科研影响力的广泛性：华东师范大学在一些热门学科（如化学、分子生物学等）有着显著的科研成果和较高的国际影响力，这表明该校的科研工作具有较强的国际竞争力，尤其在自然科学领域。

再针对文献数量做一个饼图：



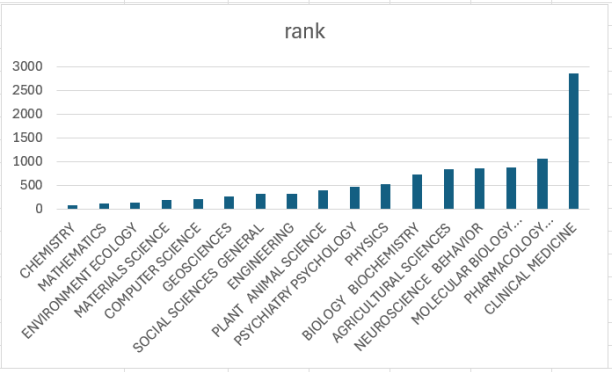
discipline	rank	docs	cites	cites_per_paper	top_papers
CHEMISTRY	90	5420	164390	30.33	157
MATHEMATICS	115	2019	1198		
ENVIRONMENT ECOLOGY	130	2941	92088		
MATERIALS SCIENCE	196	2720	93969		
COMPUTER SCIENCE	207	1803	22336		
GEOSCIENCES	275	1850	42158		
SOCIAL SCIENCES GENERAL	314	2176	27524		
ENGINEERING	317	2567	55450		
PLANT ANIMAL SCIENCE	395	1375	21843		
PSYCHIATRY PSYCHOLOGY	467	1460	15245		
PHYSICS	522	3495	50802		
BIOLOGY BIOCHEMISTRY	721	897	20837		
AGRICULTURAL SCIENCES	845	346	6513		
NEUROSCIENCE BEHAVIOR	853	771	14295		
OLECULAR BIOLOGY GENETI	867	532	20568		
HARMACOLOGY TOXICOLOG	1064	289	5693		
CLINICAL MEDICINE	2852	940	16875		

学科画像（2）：

1. 化学学科文献数量突出，5420篇 占据了较大的比例，显示出华东师范大学在化学领域有着强大的研究产出和广泛的学术影响力。这表明该学科在国际和国内的学术界具有重要的地位，且该校在该学科的研究投入与成果持续增长。
2. 生物学和分子生物学的学术影响力较强，文献数量较多，分别为 897 和 532。这表明华东师范大学在生命科学和生物医学领域的研究较为活跃，且取得了显著成果。
3. 社会科学领域文献较为均衡，达到了 2176篇，表明该学科在学术界有一定的影响力，但相较于自然科学领域，数量上有所不足。这可能表明社会科学领域仍有较大的发展潜力。
4. 较弱的学科：农业科学和药理学与毒理学的文献数量较少（分别为 346 和 289），这可能表明这些学科的学术产出相对较低，或是研究的重点尚未达到更广泛的国际认可。
5. 学科多样性：华东师范大学的学科涵盖了自然科学、生命科学和社会科学，从文献分布来看，虽然自然科学（如化学、物理、地学）占据了文献的较大比例，但生命科学和社会科学的学术贡献也不可忽视，体现了学校多学科交叉的特点。
6. 学科发展潜力：从文献数量的分布来看，化学和生物学的领先地位仍然保持，而社会科学和医学等学科的逐步提升表明该校的学科发展潜力巨大，尤其在国际化和跨学科研究方面有较大的发展空间。

最后，把各学科排名做成一个柱状图：

	A	B	C	D	E	F	G	H	I	J	K
1	discipline	rank	docs	cites	cites_per_paper	top_papers					
2	CHEMISTRY	90	5420	164390	30.33	157					
3	MATHEMATICS	115	2019	11984	5.94	22					
4	ENVIRONMENT ECOLOGY	130	2941	92088	31.31	101					
5	MATERIALS SCIENCE	196	2720	93969							
6	COMPUTER SCIENCE	207	1803	22336							
7	GEOSCIENCES	275	1850	42158							
8	SOCIAL SCIENCES GENERAL	314	2176	27524							
9	ENGINEERING	317	2567	55450							
10	PLANT ANIMAL SCIENCE	395	1375	21843							
11	PSYCHIATRY PSYCHOLOGY	467	1460	15243							
12	PHYSICS	522	3495	50802							
13	BIOLOGY BIOCHEMISTRY	721	897	20837							
14	AGRICULTURAL SCIENCES	845	346	6513							
15	NEUROSCIENCE BEHAVIOR	853	771	14295							
16	OLECULAR BIOLOGY GENETI(	867	532	20568							
17	HARMACOLOGY TOXICOLOG	1064	289	5693							
18	CLINICAL MEDICINE	2852	940	16875							
19											
20											
21											
22											
23											
24											



这个图有什么用？纵轴是很有意思的，它划分了不同排名区间，500为一组

学科画像（3）：

1. 首先，不难看出，华师大的世界前500学科是非常多的，形象来讲，前面一半以上的学科排名对应的柱子都很矮，低于500那条线
2. 500-1000 名的学科也有几个，只有最后一个是2500+的
3. **学科多样性**：华东师范大学的学科涵盖了从**自然科学**（如化学、物理、生命科学）到**社会科学**（如社会学、心理学）等多个领域。大部分学科排名集中在500名内，表现平稳，这显示出学校的学科发展较为均衡，特别是在自然科学领域。
4. 排名较高的学科（例如化学、计算机科学）显示出华东师范大学的科研优势，而排名较低的学科（如农业科学）可能需要更多的资源和投入来提升其国际影响力。
5. **排名与学术产出之间的关系**：从图表的分布来看，排名较高的学科（如化学、计算机科学）通常也伴随着较高的学术产出（文献和引用次数），而排名较低的学科（如农业科学、药理学）则在学术产出上显得较弱，显示出排名与科研影响力之间的一定关联。

最终，我们融合一下三次探索性分析得到的三个华师大的学科画像：

1. **化学学科的科研优势**

   化学是华东师范大学最强的学科之一，排名在前（90）且文献数量（5420篇）和引用次数（164390）非常突出，表明化学领域的研究成果在国际上有较高的影响力。该学科在国际学术界占据重要地位，且该校在该学科的研究投入与成果持续增长。

2. **生命科学的学术影响力**

   生物学与生物化学和**分子生物学**的文献数量也较为突出，分别为 897 和 532，显示出华东师范大学在生命科学领域的研究较为活跃，并取得显著成果。尽管这些学科的排名不如化学，但它们依然在学术界具有一定的影响力。

3. **社会科学的潜力**

   社会科学文献数量为 2176篇，表现出该学科在学术界的影响力较为均衡，但相较于自然科学领域，数量上有所不足。这可能表明社会科学领域仍有较大的发展潜力。该学科在学术产出和国际影响力上仍有进一步提升的空间。

4. **学科的相对弱势**

   农业科学和**药理学与毒理学**的文献数量较少，分别为 346 和 289，显示出这些学科的学术产出相对较低，或是这些学科的研究重点尚未达到更广泛的国际认可。需要更多的资源和学术投入，以提升它们的科研影响力。

5. **学科多样性**
   华东师范大学的学科覆盖范围广泛，涵盖了自然科学（如化学、物理、生命科学等）、生命科学（如生物学、医学等）以及社会科学（如社会学、心理学等）。大部分学科排名集中在**500名以内**，表现平稳，显示出学校学科发展较为均衡，尤其在自然科学领域。

6. **排名与学术产出之间的关系**
   从图表的分布来看，排名较高的学科（如化学、计算机科学）通常也伴随着较高的学术产出（文献和引用次数），而排名较低的学科（如农业科学、药理学）则在学术产出上显得较弱，显示出排名与科研影响力之间的一定关联。

7. **科研影响力的广泛性**
   华东师范大学在一些热门学科（如化学、分子生物学等）有着显著的科研成果和较高的国际影响力，这表明该校的科研工作具有较强的国际竞争力，尤其在自然科学领域。

8. **学科发展潜力**

   从文献数量的分布来看，**化学**和**生物学**的领先地位仍然保持，而**社会科学**和**医学**等学科的逐步提升表明该校的学科发展潜力巨大，尤其在国际化和跨学科研究方面有较大的发展空间。

10. 请利用数据建模的方式，对各学科做一个排名模型，能够较好的预测出排名位置。

(可以用各学科前60%的数据作为训练集，后20%的数据作为测试集)

数据分为三类：训练集、验证集、测试集

测试集在训练模型时，不能让模型知道，否则测试就没有意义了

首先，我们读取每个学科的表格，将输出都放到outputs文件夹中

```
CLEAN_DIR = (Path(__file__).resolve().parent / "clean")
OUT_DIR = (Path(__file__).resolve().parent / "outputs")
RANDOM_STATE = 42
os.makedirs(OUT_DIR, exist_ok=True)
```

然后搭建脚本的主循环框架：

先用 glob 找到 clean 目录下的所有 CSV 文件并按文件名排序，然后按顺序逐个打开处理

对每个文件读取时用 try/except 包裹以防单个文件损坏或格式异常导致整个批处理终止

一旦成功读取，就打印学科名与数据的行列信息，作为处理的上下文说明

```
csvs = sorted(glob.glob(os.path.join(CLEAN_DIR, "*.csv")))
if not csvs:
    print("未在 'clean' 目录找到任何 csv 文件。")
for f in csvs:
    try:
        df = pd.read_csv(f)
    except Exception as e:
        print(f"读取失败：{f}，跳过。错误：{e}")
        continue

    subject = os.path.splitext(os.path.basename(f))[0]
    print(f"\n处理学科：{subject}，行数={len(df)}，列数={len(df.columns)}")
```

我们可以直接把表格的第一列当作排名

机构名称优先取 Institutions 列，如果那列不存在则退回取第一个字符串类型的列

如果连字符串列都没有，则假定第二列是机构名并用它

```
y = pd.to_numeric(df.iloc[:, 0], errors='coerce')  # 第一列为排名，强制转数值
# Institutions 列优先，否则第一个 object 列，否则退回第二列
if 'Institutions' in df.columns:
    names = df['Institutions'].astype(str)
else:
    obj_cols = [c for c in df.columns if df[c].dtype == object]
    names = df[obj_cols[0]].astype(str) if obj_cols else df.iloc[:,
1].astype(str)
```

把除排名和机构名以外的列当作特征 X

对于**非数值**列我们采用 pd.get_dummies 做 one-hot 编码

对于**数值**列则用中位数填补缺失值

最后，确保 X 的所有列都转换为浮点数以便后续模型可以直接接受

```
drop_cols = [df.columns[0]]
if 'Institutions' in df.columns:
    drop_cols.append('Institutions')
X = df.drop(columns=drop_cols, errors='ignore').copy()

nonnum = X.select_dtypes(exclude=[np.number]).columns.tolist()
if nonnum:
    X = pd.get_dummies(X, columns=nonnum, dummy_na=True, drop_first=True)

for c in X.columns:
    if X[c].isna().any():
        col_med = X[c].median()
        X[c].fillna(col_med if not np.isnan(col_med) else 0, inplace=True)

X = X.astype(float)
```

如何划分数据?

首先把整体数据的 20% 随机切为测试集，剩下的 80% 再从中按 25% 切出验证集

这样得到最终的 60% 训练、20% 验证、20% 测试

```
# 6. split: first test 20%, then from remaining split 25% as val (-> total
60/20/20)
X_temp, X_test, y_temp, y_test, names_temp, names_test = train_test_split(
    X, y, names, test_size=0.2, random_state=RANDOM_STATE, shuffle=True
)
X_train, X_val, y_train, y_val, names_train, names_val = train_test_split(
    X_temp, y_temp, names_temp, test_size=0.25, random_state=RANDOM_STATE,
shuffle=True
)
```

然后进行简单的验证调参，选择 best_depth

用验证集在两个 5 与 None 之间选择哪个能得到更低的验证集 MAE

```
best_depth = None
best_mae = 1e9
for depth in (5, None):
    model = RandomForestRegressor(n_estimators=200, max_depth=depth,
random_state=RANDOM_STATE, n_jobs=-1)
    model.fit(X_train, y_train)
    p = model.predict(X_val)
    m = mean_absolute_error(y_val, p)
    if m < best_mae:
        best_mae = m
        best_depth = depth
```

最后，我们把训练集和验证集合并以训练最终模型

训练时我们把树的数量 n_estimators 设为 300 以取得更平滑的预测

然后在测试集上得到原始浮点预测值 pred_raw 并同时把它四舍五入为整数 pred_round

计算测试集上的 MAE 并把 Institutions、真实排名、四舍五入预测排名与原始预测分数保存为一个 CSV 文件

最后在屏幕上打印测试集的一部分表格和 MAE 值，以便实时观察情况

```
X_comb = pd.concat([X_train, X_val], ignore_index=True)
y_comb = pd.concat([y_train, y_val], ignore_index=True)
final = RandomForestRegressor(n_estimators=300, max_depth=best_depth,
random_state=RANDOM_STATE, n_jobs=-1)
final.fit(X_comb, y_comb)

pred_raw = final.predict(X_test)
pred_round = np.rint(pred_raw).astype(int)
mae = mean_absolute_error(y_test, pred_raw)

out = pd.DataFrame({
    'Institutions': names_test,
    'true_rank': y_test.values,
    'pred_rank': pred_round,
    'pred_rank_raw': pred_raw
})
out = out.sort_values('true_rank').reset_index(drop=True)

out_path = os.path.join(OUT_DIR, f"{subject}_test_predictions.csv")
out.to_csv(out_path, index=False, encoding='utf-8-sig')
print(out[['Institutions', 'true_rank', 'pred_rank']])
print(f"MAE (test) = {mae:.4f}  -> 已保存: {out_path}")
```

完整代码见 10_machine_learning.py

测试结果保存到了outputs文件夹中

我还打印了一份总表，对于每个学科，给出样本数量和测试集上的平均绝对误差

	A	B	C	D	E
1	subject	n_samples	mae_test		
2	AGRICULTURAL SCIENCES	1381	3.5559728		
3	BIOLOGY & BIOCHEMISTRY	1649	5.3479711		
4	CHEMISTRY	2141	7.6695684		
5	CLINICAL MEDICINE	6754	9.9666308		
6	COMPUTER SCIENCE	863	5.0602249		
7	ECONOMICS & BUSINESS	543	6.0048879		
8	ENGINEERING	2787	7.1374907		
9	ENVIRONMENT ECOLOGY	2066	4.990157		
10	GEOSCIENCES	1175	7.3136851		
11	IMMUNOLOGY	1177	3.4091789		
12	MATERIALS SCIENCE	1580	6.6232756		
13	MATHEMATICS	395	8.9823825		
14	MICROBIOLOGY	803	3.1964711		
15	MOLECULAR BIOLOGY & GENETICS	1169	6.252569		
16	MULTIDISCIPLINARY	216	1.3183143		
17	NEUROSCIENCE & BEHAVIOR	1298	5.2798264		
18	PHARMACOLOGY & TOXICOLOGY	1389	3.1396763		
19	PHYSICS	995	13.80468		
20	PLANT & ANIMAL SCIENCE	1950	5.0129139		
21	PSYCHIATRY PSYCHOLOGY	1147	5.4386532		
22	SOCIAL SCIENCES, GENERAL	2407	4.6220652		
23	SPACE SCIENCE	236	18.91298		
24					
25					
26					