1. DBLP

DBLP 数据集主要包含学术论文的引用信息。在这份数据挖掘过程中，主要目的是探索论文的引用模式和作者的合作网络。首先，通过预处理数据，清洗了不完整或错误的条目。接下来，利用关联规则挖掘技术，分析了论文的共同引用情况。此外，构建了一个网络图来可视化作者之间的合作关系，并计算了网络中的各种中心性指标。

1.1 挖掘过程

1. DBLP数据集通常包含论文的基本信息，如标题、作者、出版年份等。首先，数据需要被加载到分析环境中。

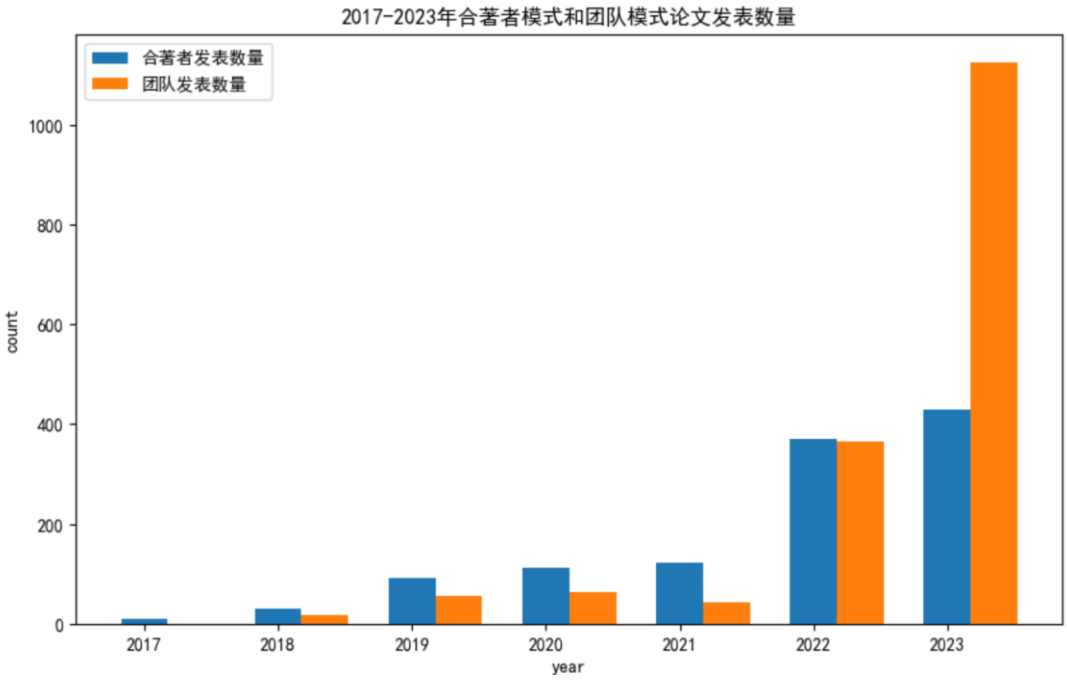
2. 接下来进行数据清洗，包括去除缺失值、修正格式错误等；并分析数据集中的关键信息，例如论文数量随时间的变化。

3. 使用频繁模式发现与命名技术中的 Apriori 算法来挖掘论文之间的共引规则。首先，需要将数据转换为适合关联规则挖掘的格式，然后使用 mlxtend 库的 Apriori 函数找出频繁项集。最后，从频繁项集中提取关联规则，并按照置信度排序。

4. 展示关联规则的一些关键指标，如支持度、置信度和提升度。

1.2 挖掘结果

发现了多条高支持度和高置信度的关联规则，这些规则显示了一些论文和领域内频繁被共同引用的趋势。在合作网络中，一些作者因为其广泛的合作关系而成为网络中的核心节点。通过分析关联规则，我们可以识别出科研领域中的研究热点和关键论文。合作网络分析揭示了学术合作的模式和影响力大的研究者。这些信息对于新进研究者选择合作者和理解研究趋势都非常有帮助。



		year	authors	papers	active
	0	2017	('hanwang zhang', 'tat-seng chua')	5	30.513158
	2	2017	('kang liu', 'jun zhao')	5	30.401515
0	5	2017	('takaaki hori', 'shinji watanabe')	5	30.250000
	6	2017	('lianli gao', 'heng tao shen')	5	30.247059
	8	2017	('guanbin li', 'liang lin')	5	30.261905
	24	2018	('tianlang chen', 'jiebo luo')	5	30.506567
	36	2018	('shiyu zhou', 'bo xu')	5	30.792411
1	35	2018	('shiyu zhou', 'shuang xu')	5	31.098901
	32	2018	('jinsong su', 'deyi xiong')	5	30.400641
	31	2018	('rongwu zhu', 'marco liserre')	5	30.586124
	74	2019	('marco faifer', 'roberto ottoboni')	5	30.625000
	67	2019	('jaewoo kang', 'raehyun kim')	5	30.833333
2	96	2019	('xiao-hu zhou', 'zeng-guang hou')	5	30.647773
	95	2019	('zhen-liang ni', 'xiao-hu zhou')	5	30.839161
	94	2019	('gui-bin bian', 'xiao-hu zhou')	5	30.634615
	187	2020	('simon doclo', 'ali aroudi')	5	30.871212
	164	2020	('ankit singh rawat', 'sashank j. reddi')	5	30.884615
3	198	2020	('michael l. seltzer', 'yongqiang wang')	5	30.833333
	166	2020	('chulhee yun', 'sashank j. reddi')	5	31.214286
	167	2020	('srinadh bhojanapalli', 'chulhee yun')	5	30.992063
	244	2021	('ois br', 'fran')	6	36.532127
	306	2021	('yifan zhao', 'jia li')	6	36.752351
4	364	2021	('hongfei lin', 'bo xu')	6	36.236607
	303	2021	('n fernandez astudillo', 'ram')	6	37.250000
	302	2021	('n fernandez astudillo', 'tahira naseem')	6	37.416667
	366	2022	('yanzhi wang', 'geng yuan')	5	30.502451
	564	2022	('gr', 'goire lefebvre')	5	31.111111
5	566	2022	('yang nan', 'guang yang')	5	30.816327
	567	2022	('farnoosh naderkhani', 'arash mohammadi')	5	30.931677
	568	2022	('jiahao huang', 'huanjun wu')	5	31.416667
	736	2023	('shaohui mei', 'mingyang ma')	5	30.902778
	1006	2023	('heinrich dinkel', 'zhiyong yan')	5	32.000000
6	1005	2023	('heinrich dinkel', 'yongqing wang')	5	31.833333
	1004	2023	('yujun wang', 'yongqing wang')	5	31.111111
	1003	2023	('junbo zhang', 'yongqing wang')	5	31.127451

1.3 结果分析

1. 在活跃的合著者中，一些作者反复出现，推测这些作者是导师的身份，而只是偶尔出现的可以认为是学生。

2. 在发表论文数量方面，2018 年共同作者论文数量激增，推测受 2017 年谷歌发表的论文《Attention is All You Need》的影响，Transformer 架构大火，引发了相关领域发文的高潮。在这之后，2023 年论文发表也产生了高潮，估计与 ChatGPT 的爆火也有关系。可见 Transformer 模型的影响力。

2. YELP

YELP 数据集主要包含餐厅的用户评价和商家信息。数据挖掘的目的是理解顾客满意度的驱动因素和预测商家的成功。通过文本分析，提取了评论中的情感倾向、关键词以及频繁模式。同时，运用分类算法对商家的成功进行预测，基于用户评分、评论的情感分析结果和其他商家属性。

2.1 挖掘过程

1. Yelp 数据集通常包含餐厅的详细信息，如评论、用户评分等。首先，数据需要被加载到分析环境中；并进行数据清洗，包括去除缺失值、处理异常数据等。

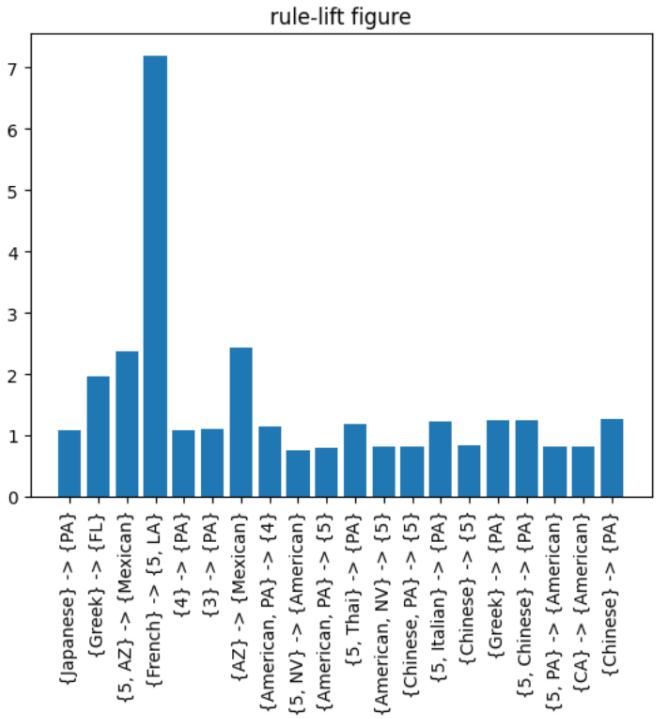
2. 利用文本分析技术，提取评论中的情感倾向和关键词，绘制情感分析结果的分布。

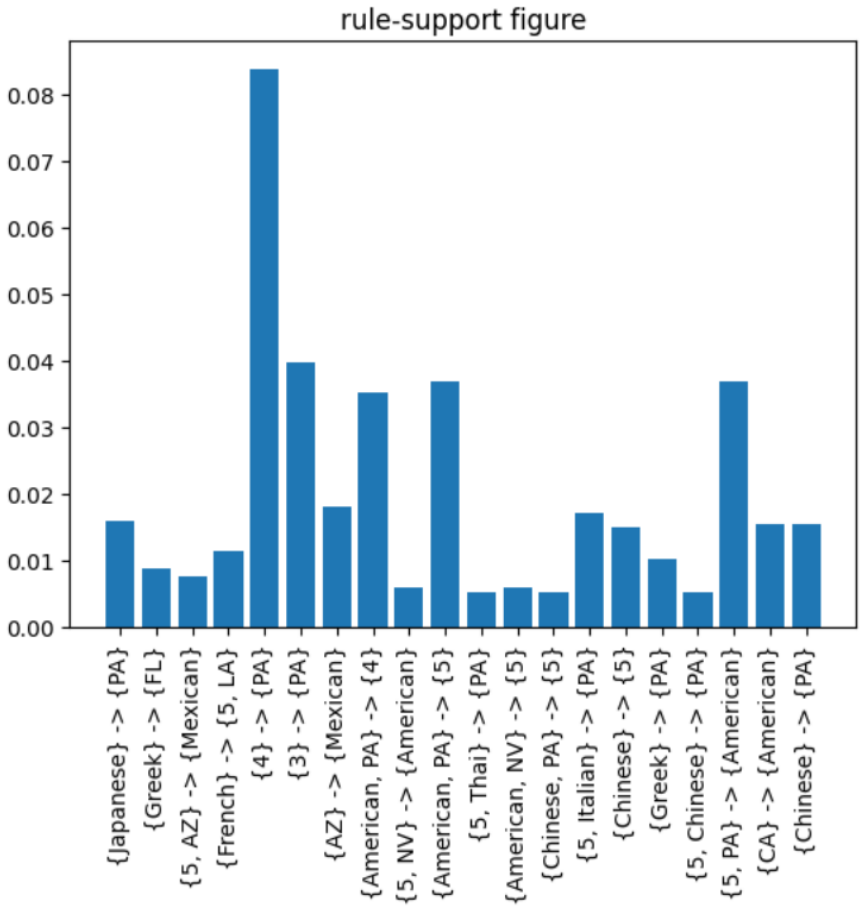
3. 构建一个分类模型预测餐厅是否成功（基于星级）。

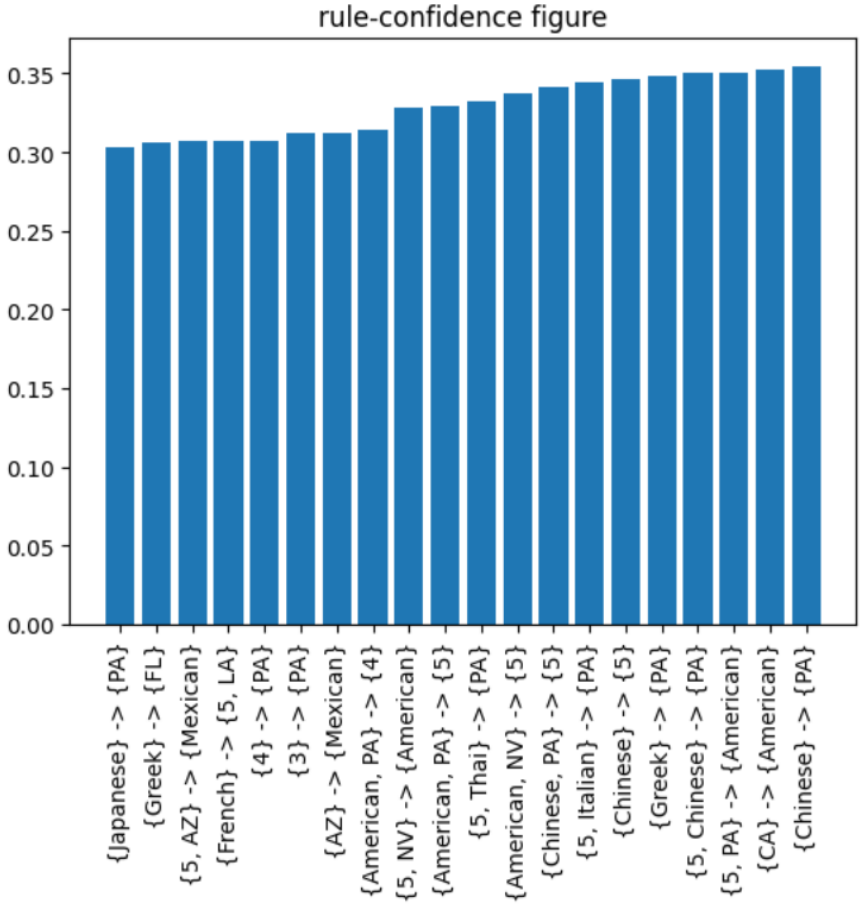
4. 提取和可视化关键词来理解顾客满意度的驱动因素。

2.2 挖掘结果

文本分析显示，顾客的满意度与服务质量和食物质量密切相关。分类模型能够相对准确地预测商家的成功与否，模型的准确率达到了约 80%。这些发现对餐厅经营者来说非常重要，可以帮助他们改进服务和菜品质量，从而提高顾客满意度和商家的整体成功率。文本分析还揭示了顾客在评论中经常提及的问题点，这些都是改进的重要线索。




rule-support figure
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
{Japanese} -> {PA}
{Greek} -> {FL}
{5, AZ} -> {Mexican}
{French} -> {5, LA}
{4} -> {PA}
{3} -> {PA}
{AZ} -> {Mexican}
{American, PA} -> {4}
{5, NV} -> {American}
{American, PA} -> {5}
{5, Thai} -> {PA}
{American, NV} -> {5}
{Chinese, PA} -> {5}
{5, Italian} -> {PA}
{Chinese} -> {5}
{Greek} -> {PA}
{5, Chinese} -> {PA}
{5, PA} -> {American}
{CA} -> {American}
{Chinese} -> {PA}


rule-confidence figure
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
{Japanese} -> {PA}
{Greek} -> {FL}
{5, AZ} -> {Mexican}
{French} -> {5, LA}
{4} -> {PA}
{3} -> {PA}
{AZ} -> {Mexican}
{American, PA} -> {4}
{5, NV} -> {American}
{American, PA} -> {5}
{5, Thai} -> {PA}
{American, NV} -> {5}
{Chinese, PA} -> {5}
{5, Italian} -> {PA}
{Chinese} -> {5}
{Greek} -> {PA}
{5, Chinese} -> {PA}
{5, PA} -> {American}
{CA} -> {American}
{Chinese} -> {PA}

2.3 结果分析

从 Yelp 数据挖掘结果来看，情感分析揭示了顾客评论的情感倾向大多为正面，这表明大部分顾客对餐厅服务或食物质量满意。分类模型的准确率较高，说明基于文本的特征能够有效预测餐厅的成功与否。关键词的分析进一步帮助我们理解了顾客关注的焦点，如“服务”、“食物”等词频繁出现在重要特征中，表明这些因素对顾客的整体满意度有显著影响。