**西安电子科技大学**

**信息计量学实验 课程实验报告**

经济与管理 学院 班

成 绩

姓名 学号

姓名 学号

实验日期 年 月 日

|  |
| --- |
| 指导教师评语：  指导教师：  年 月 日 |
| **实验报告内容基本要求及参考格式**  一、实验目的  二、实验所用仪器（或实验环境）  三、实验基本原理及步骤（或方案设计及理论计算）  四、实验数据记录（或仿真及软件设计）  五、实验结果分析及回答问题（或测试环境及测试结果） |

**这里可以自拟标题**

**——基于文献数据库的信息计量研究**

## 1、选题及数据准备

1.1选题理由

Large language model

选择"Large language model"作为基于文献数据库的信息计量研究主题有以下理由：

1）新兴领域：Large language models，如GPT-3.5，代表了自然语言处理和人工智能领域的最新进展。它们具有强大的语言理解和生成能力，可以用于各种应用领域，包括文本生成、对话系统、机器翻译等。由于其重要性和广泛应用，研究大型语言模型的信息计量是有意义的。

2）影响力和研究产出：大型语言模型如GPT-3.5在自然语言处理和人工智能领域引起了广泛的关注和讨论。研究关于大型语言模型的文献数量和引用数量可能反映了该领域的发展和影响力。通过信息计量研究，可以分析大型语言模型相关研究的趋势、研究合作网络、研究贡献等，以进一步了解该领域的研究动态。

3）信息检索和知识获取：大型语言模型具有强大的文本生成和问答能力，可以用于信息检索和知识获取。通过研究大型语言模型在文献数据库中的应用，可以探索如何利用这些模型提高文献检索的效果，从海量文献中获取有用的知识和信息。

4）方法论和技术创新：研究大型语言模型在文献数据库中的信息计量方法和技术也具有创新意义。如何利用大型语言模型进行文献的自动分类、主题建模、情感分析等，可以为文献数据库的信息管理和分析提供新的思路和方法。

综上所述，选择"Large language model"作为基于文献数据库的信息计量研究主题是合理的，因为它代表了新兴领域的研究进展，具有影响力和研究产出，有助于信息检索和知识获取，并且可以推动方法论和技术创新。

1.2数据检索

1. 选择数据库：我们选择使用Web of science和SCOPUS数据库进行数据检索。

对比：

在CNKI中搜索大语言模型：

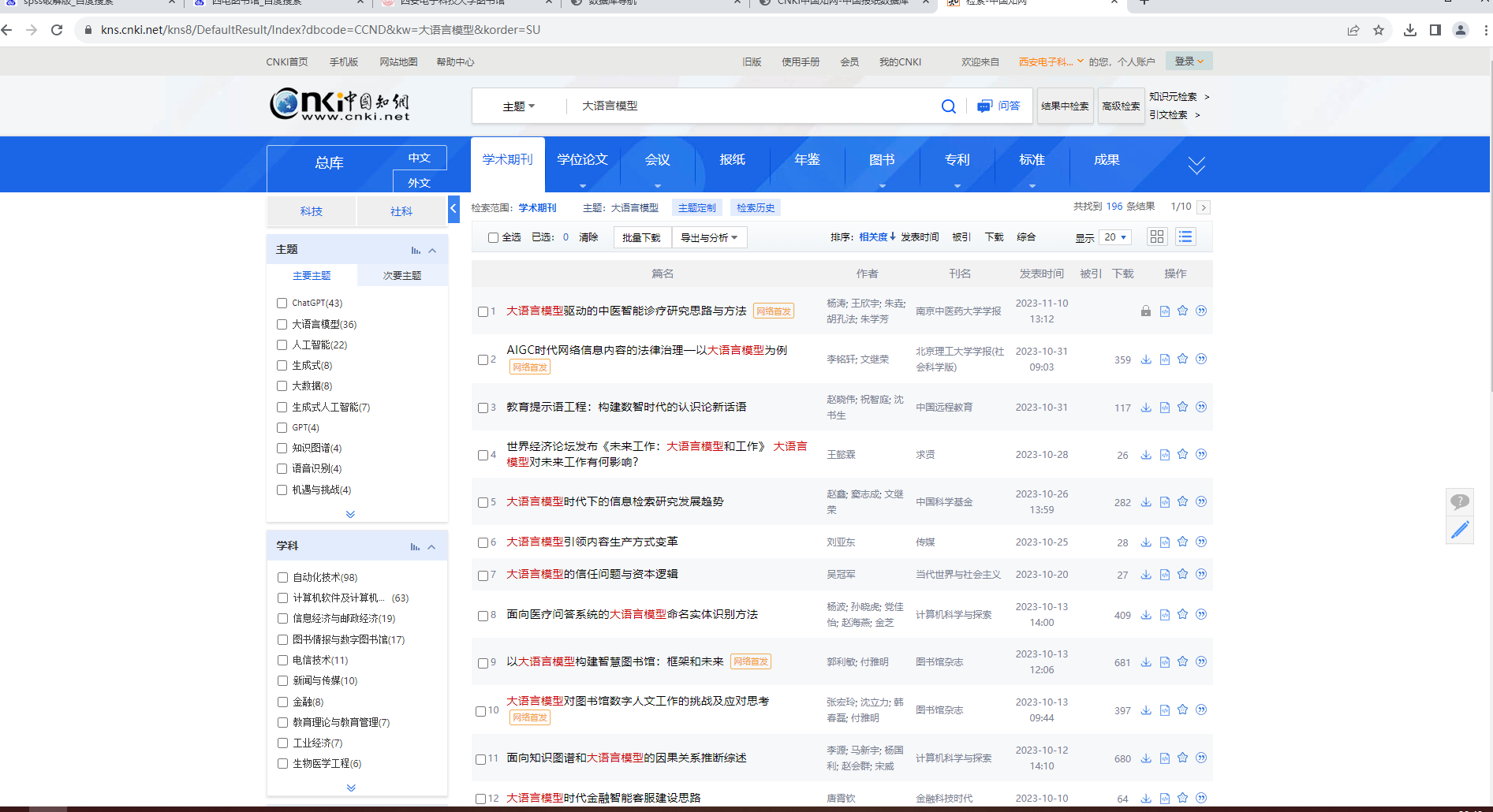


图1.1 CNKI搜索

显示结果仅有196条。

在CSSCI中搜索大语言模型：



图1.2 CSSCI搜索

显示结果仅有2条。

在Web of science中搜索Large language model：

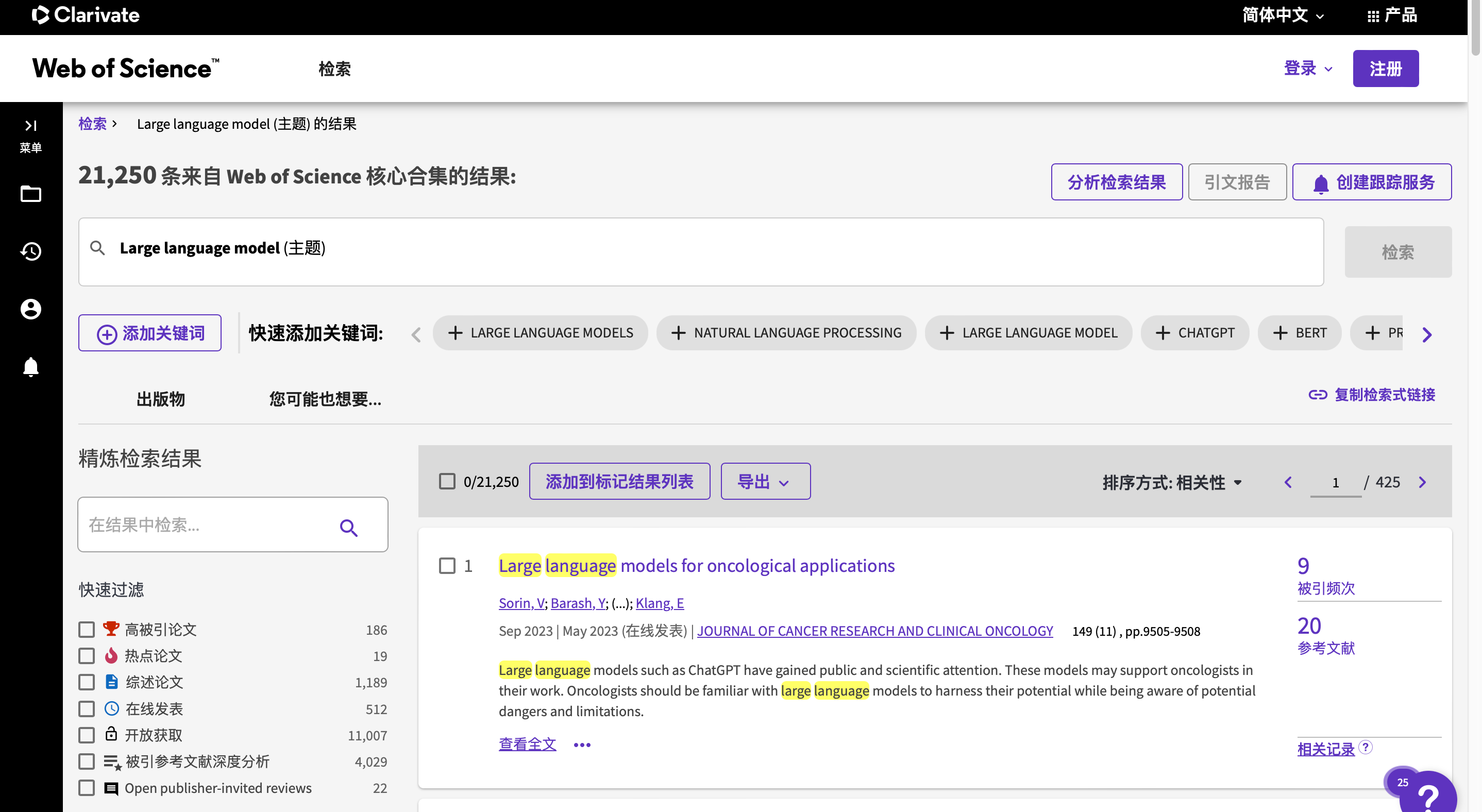


图1.3 Web of science搜索

显示有21250条结果。

在SCOPUS中搜索Large language model：



图1.4 SCOPUS搜索

显示有41822条结果。

故选择使用Web of science和SCOPUS数据库进行数据检索。

2）确定关键词：根据研究主题，确定Large language model关键词来描述。

3）执行检索：在wos数据库中执行构建好的检索查询，并获取检索结果。

4）调整检索策略：根据检索结果进行调整和优化检索策略。添加限制条件的方法，如在开始检索前选择web of science核心合集内检索，文献类型选择论文，研究方向选择计算机科学，类别选择计算机科学人工智能等。

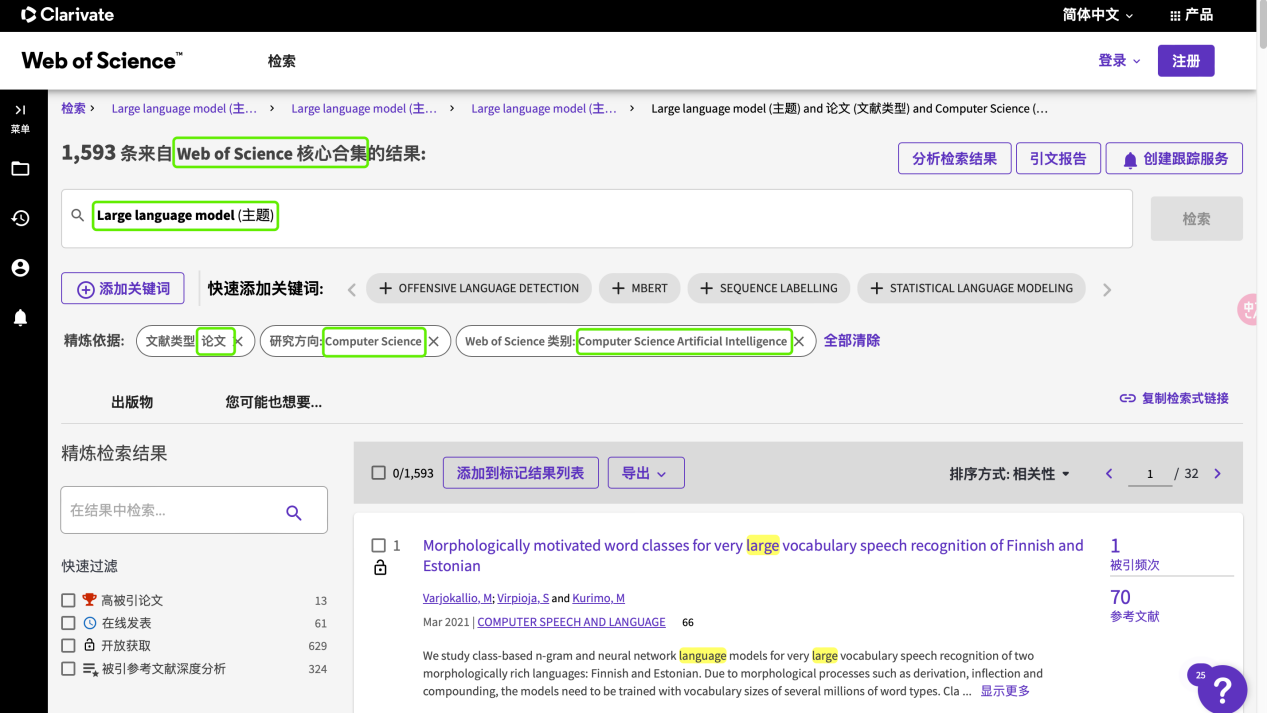


图1.5 web of science优化检索

5）报告数据检索结果：根据检索结果进行数据分析和报告。可以统计检索到的文献数量、分析文献的作者、机构、发表时间等信息，以及文献的引用情况等。

1.3数据清洗

数据清洗包括剔除没有作者的文章、去除重复文献、处理缺失值等。

1）剔除没有作者的文章：可以通过筛选和过滤操作，将没有作者信息的文献从数据集中移除。

2）去除重复文献：通过比较文献的标题、作者、摘要等信息，识别和去除重复的文献记录。

3）处理缺失值：如果数据集中存在缺失值，选择删除包含缺失值的文献记录。

清洗后的数据集将更加干净和可靠，有助于后续的数据分析和研究结果的准确性。

1.4报告数据集的概况

对于基于文献数据库的信息计量研究，数据集的概况包括以下方面的信息：

1）数据来源：Web of Science数据库。

2）文献数量：1593个。

3）时间范围：指定数据集中文献的时间范围，例如从哪一年到哪一年的文献被包括在数据集中。

4）文献类型：期刊论文。

5）领域覆盖：计算机科学人工智能。

6）引用信息：数据集中的文献具有引用信息，报告平均引用数量、最高引用数量等。

7）数据清洗：如果进行了数据清洗过程，可以简要描述清洗的步骤和策略。

## 2、传统信息计量规律的验证

根据scopus数据库进行相关操作：

2.1验证信息增长规律的步骤：

1）统计每年有关Large language model主题的文献数量。

筛选中排除2024的数据，限定研究范围是计算机科学，文献类型是论文。

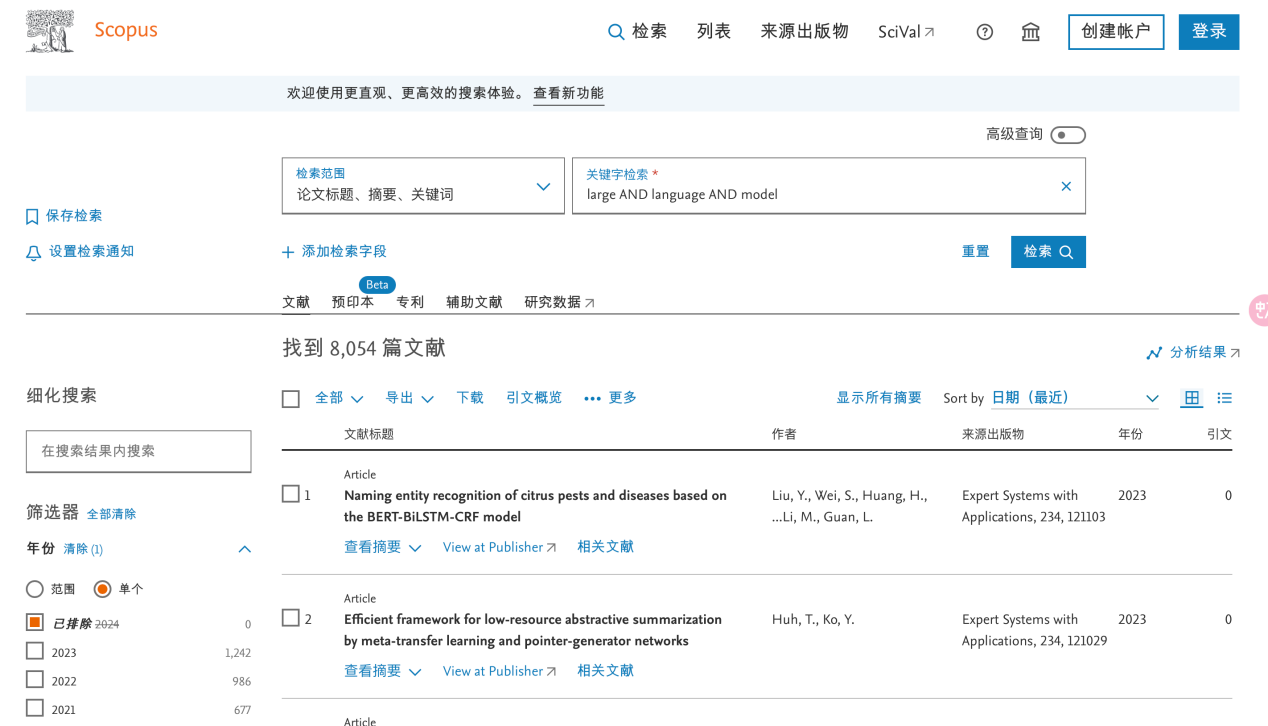


图2.1 scopus检索分析

2）观察每年的增长变化趋势。

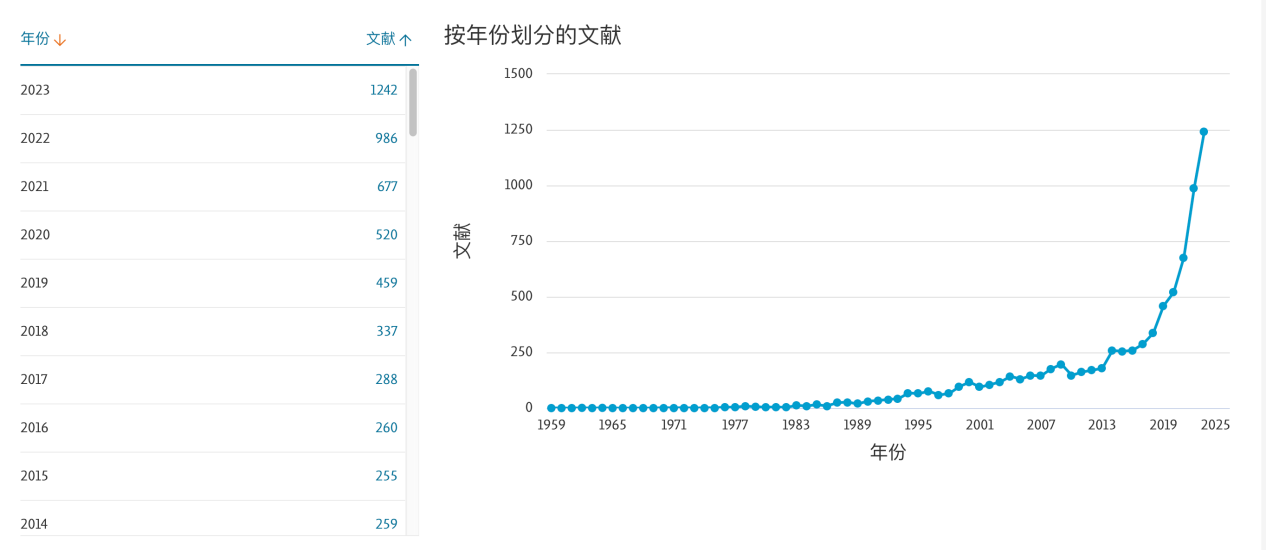


图2.2 scopus总年份变化

3）分析可知2000年后文献数量开始显著变化，可以选择00年后的数据进行分析。

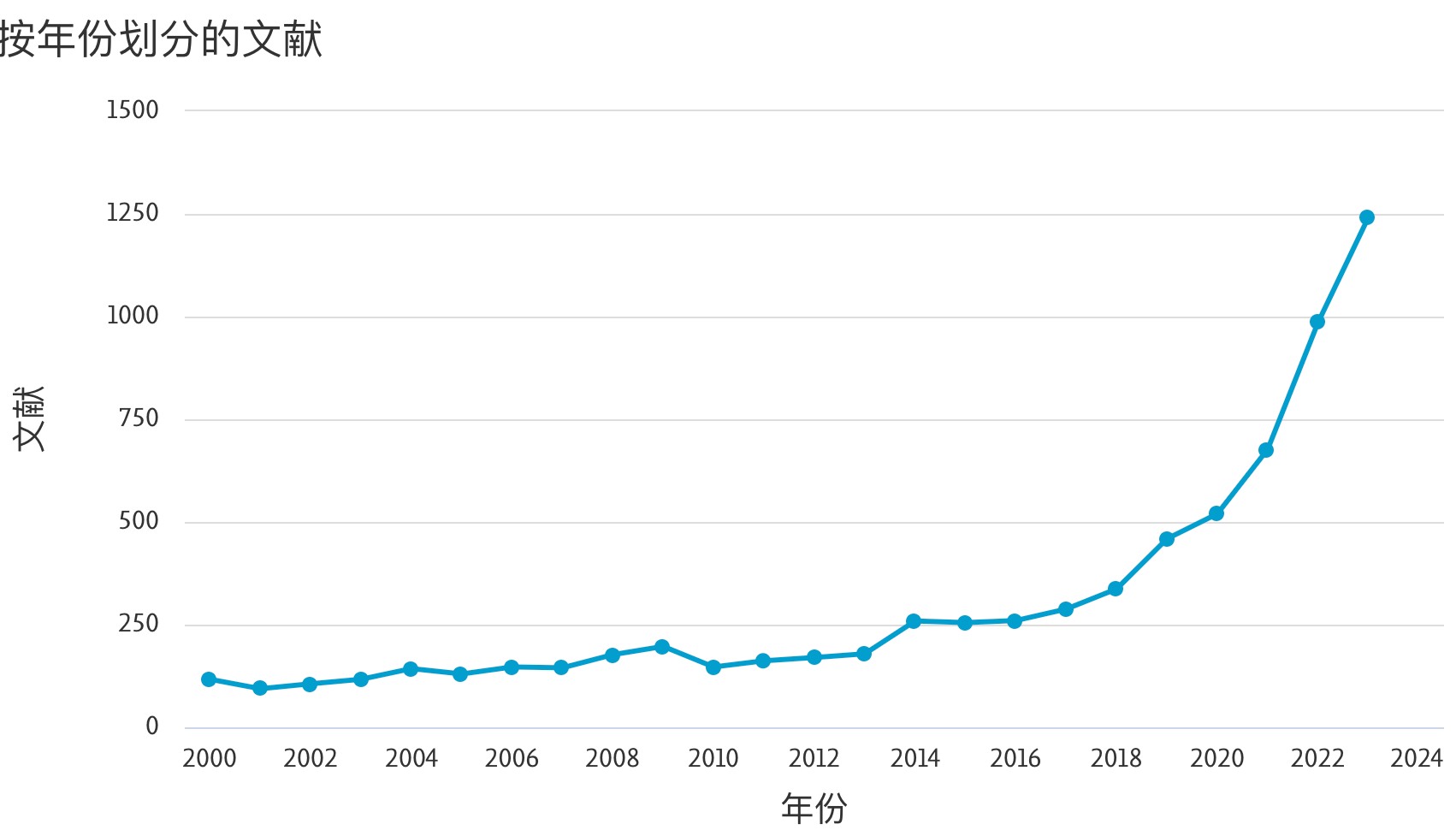


图2.3 筛选后年份变化图

4）文献数据如下：



图2.4 验证增长的数据

5）拟合：

假设模型为：

F（t）=a\*e^bt

a：统计初识时刻的文献量

b：持续增长率

t：时间，年为单位

根据以上数据，将时间数据变为从0-23，a=117，使用python拟合这些数据点到指数函数得到最终的拟合结果为：

b =0.14753739873311938=0.1475

即F（t）=117\*e^0.1475t

参数b的正值表明累积数量以指数形式增长，增长速率较快。每增加一年，累积数量增加的比例是指数级增长的。

6）结论：

综上所述，大语言模型领域的文献增长符合信息增长规律，目前有关大语言模型的研究正处于“加速——饱和——更大加速”中的加速增长阶段，呈现阶梯型指数增长，且增长速率较快。这反映了该领域的研究兴趣和活跃度的增加，以及科学研究的持续发展和扩张。

2.2验证布拉德福定律：集中和分散广泛存在

1）研究方法：选择区域法。

2）研究对象：“large language model”为主题的相关论文和对于的期刊进行研究。

3）数据来源：

选取 Cnki 数据库为数据源，对“large language model”为主题进行搜索，时间设定为 2000 年到 2023 年，共搜索到人工智能相关论文7314篇。对期刊进行倒叙排序并选取论文数在前160个的期刊进行数据可视化分析，以中所含相关论文数最多。

得到以下数据：

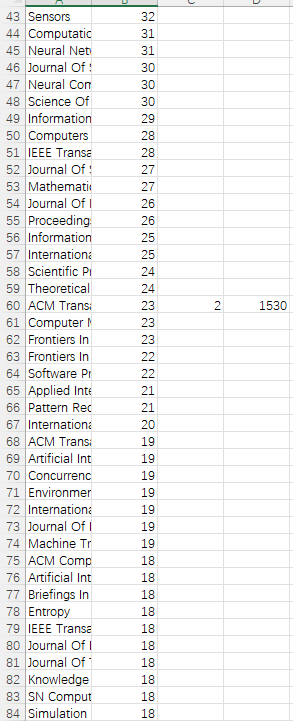
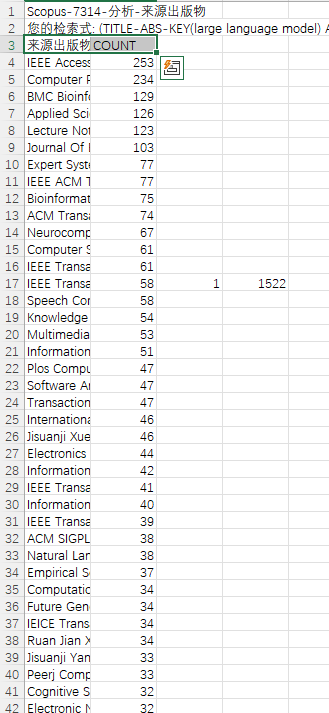


图2.5.1 验证布拉德福定律的数据

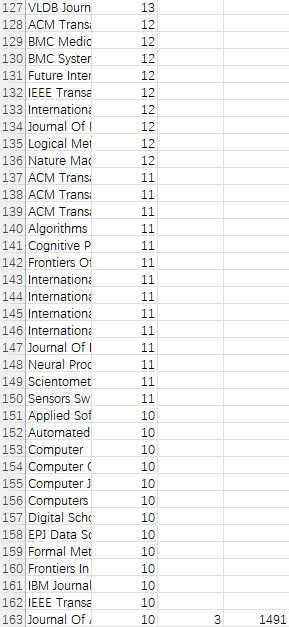
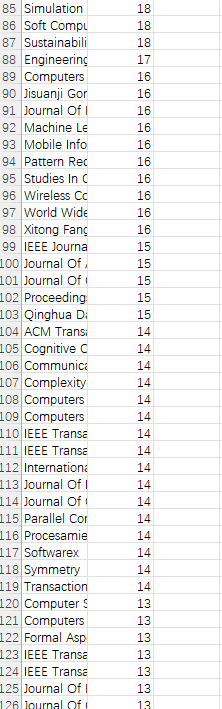


图2.5.2 验证布拉德福定律的数据

4）以排名前十的期刊作为例子，做出折线可视化

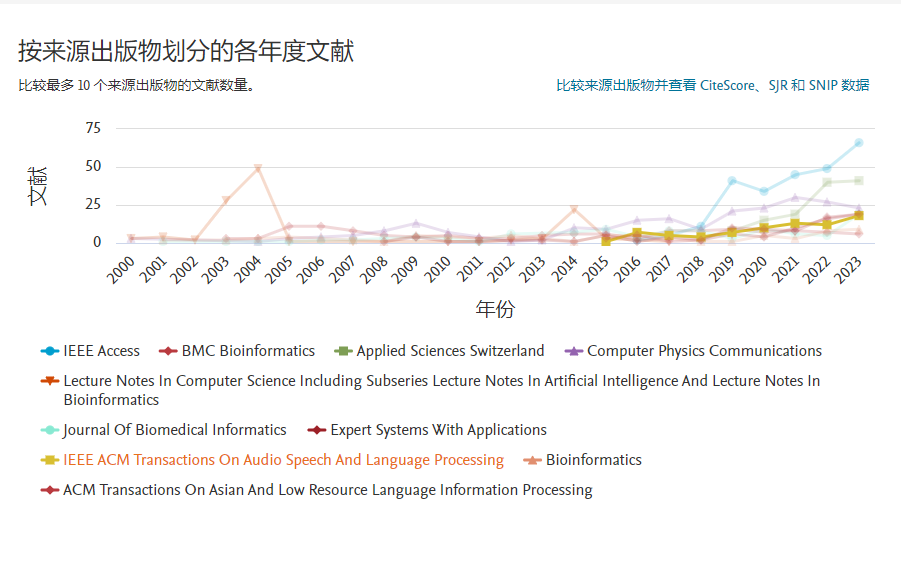


图2.6 文献数据可视化

5）数据分析：

为观察实际值与理论值的差异，下面我们利用拟合公式对数据进行比较分析。按照布 拉德福区域分析法的描述，将某一时段的期刊，按某一学科的载文量递减顺序排列后，分为三个区，令每一个区内包含的论文数相等，即每个区大约含有1530篇论文。



图2.7 文献分区

根据对三分区的拟合中发现，实验数据具有相似的布氏定律分布特征，n=3，核心区：相关区：非相关区=1: 3：9

6）结论：

这一现象说明：主题为“large language model”的文章分布期刊符合布拉德福定律，在这种情况下，我们可以推断该领域的研究已经形成了一定的学术共同体，研究人员在选择期刊发表文章时，会遵循一定的规则和惯例。这有助于确保该领域的期刊分布具有一定的连续性和稳定性，从而促进学术交流和知识传播。

此外，这也可能意味着该领域的研究已经相对成熟，有稳定的期刊出版渠道和评审机制。同时，布拉德福定律的符合也可能表明该领域的期刊分布具有一定的学科交叉性，因为布拉德福定律通常被认为是一种跨学科现象。

总之，主题为“large language model”的文章分布期刊符合布拉德福定律，说明了该领域期刊分布的规律性、稳定性、成熟度和学科交叉性等特点。

## 3、引文分析

1）选择Hiscite软件

2）获得引文数据：

在wos数据库中导出数据。

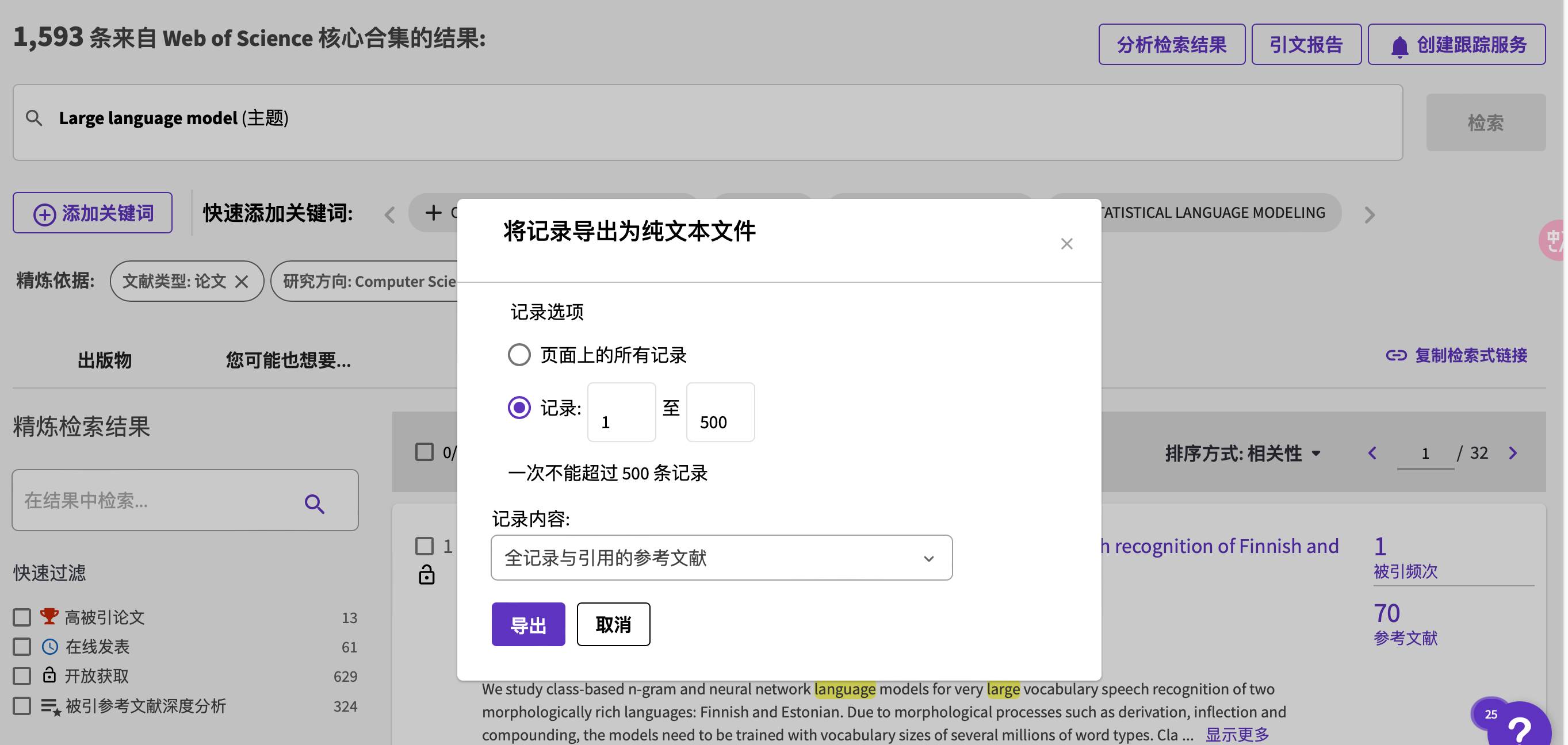


图3.1 导出数据

分四次导出1593条的数据。

3）导出引文数据：

将 Science 换成 Knowledge

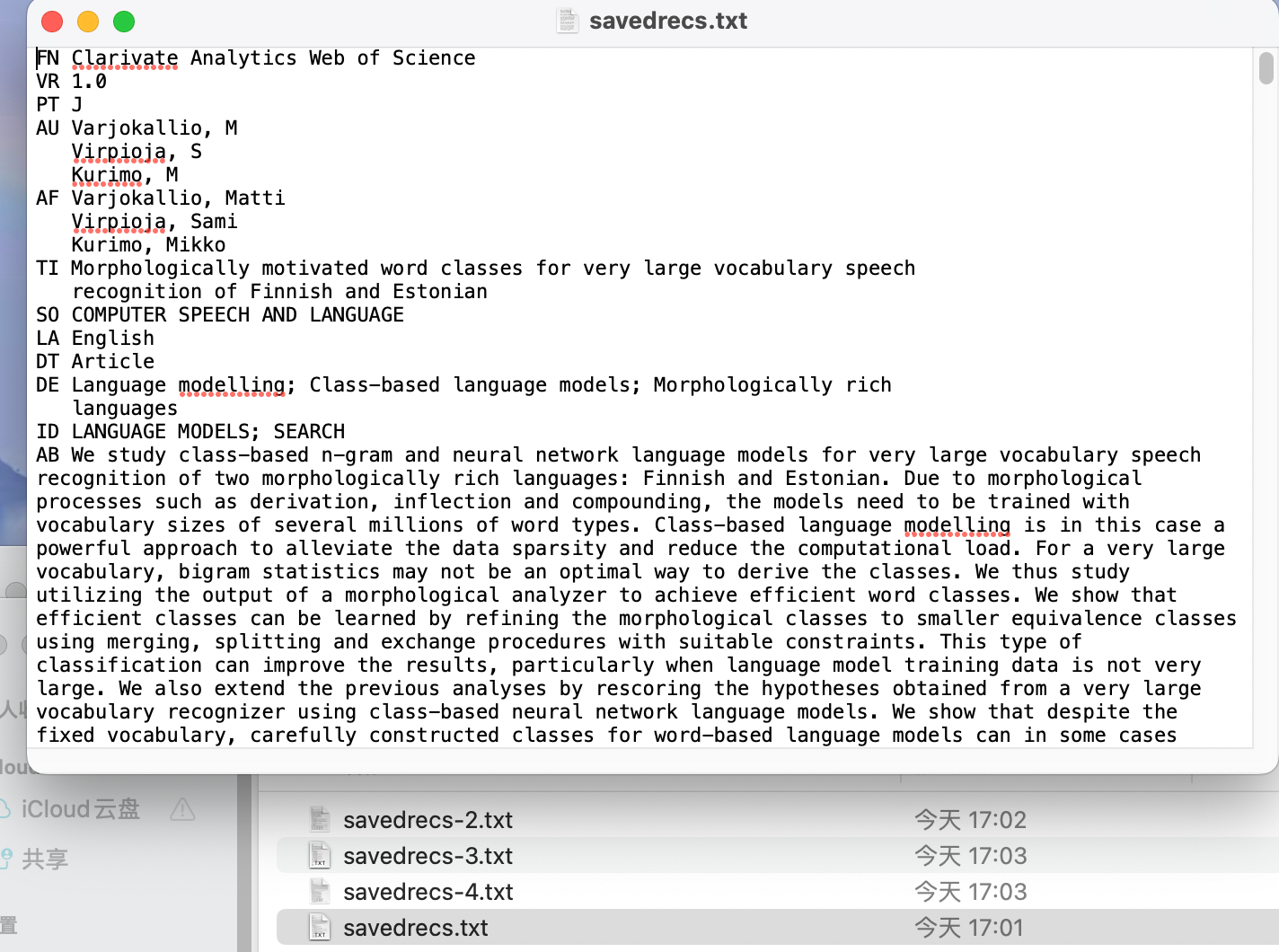


图3.2 修改前

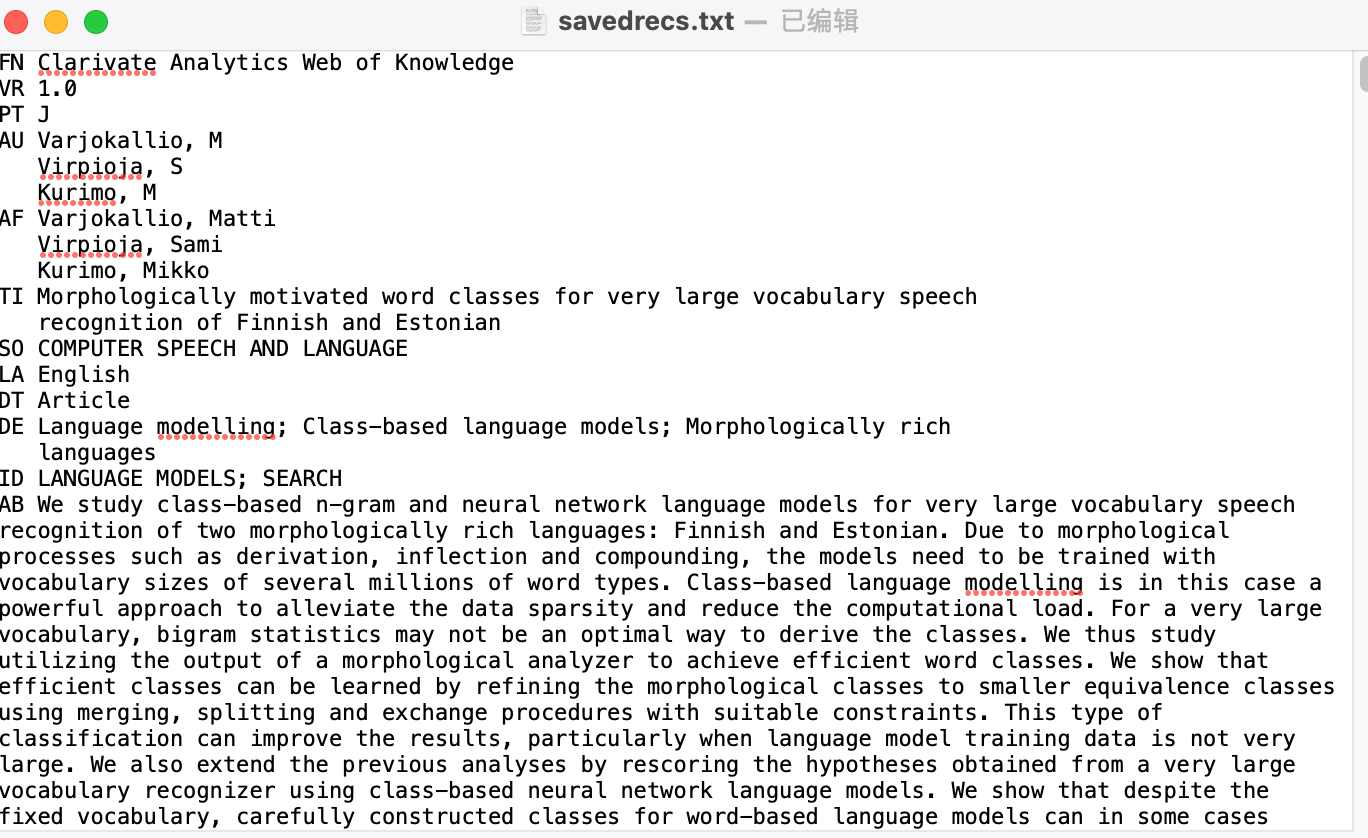


图3.3 修改后

1. 导入数据：

运行 Histcite 软件，依次将多个文本文件添加入 HisiCite。

导入成功，如图。

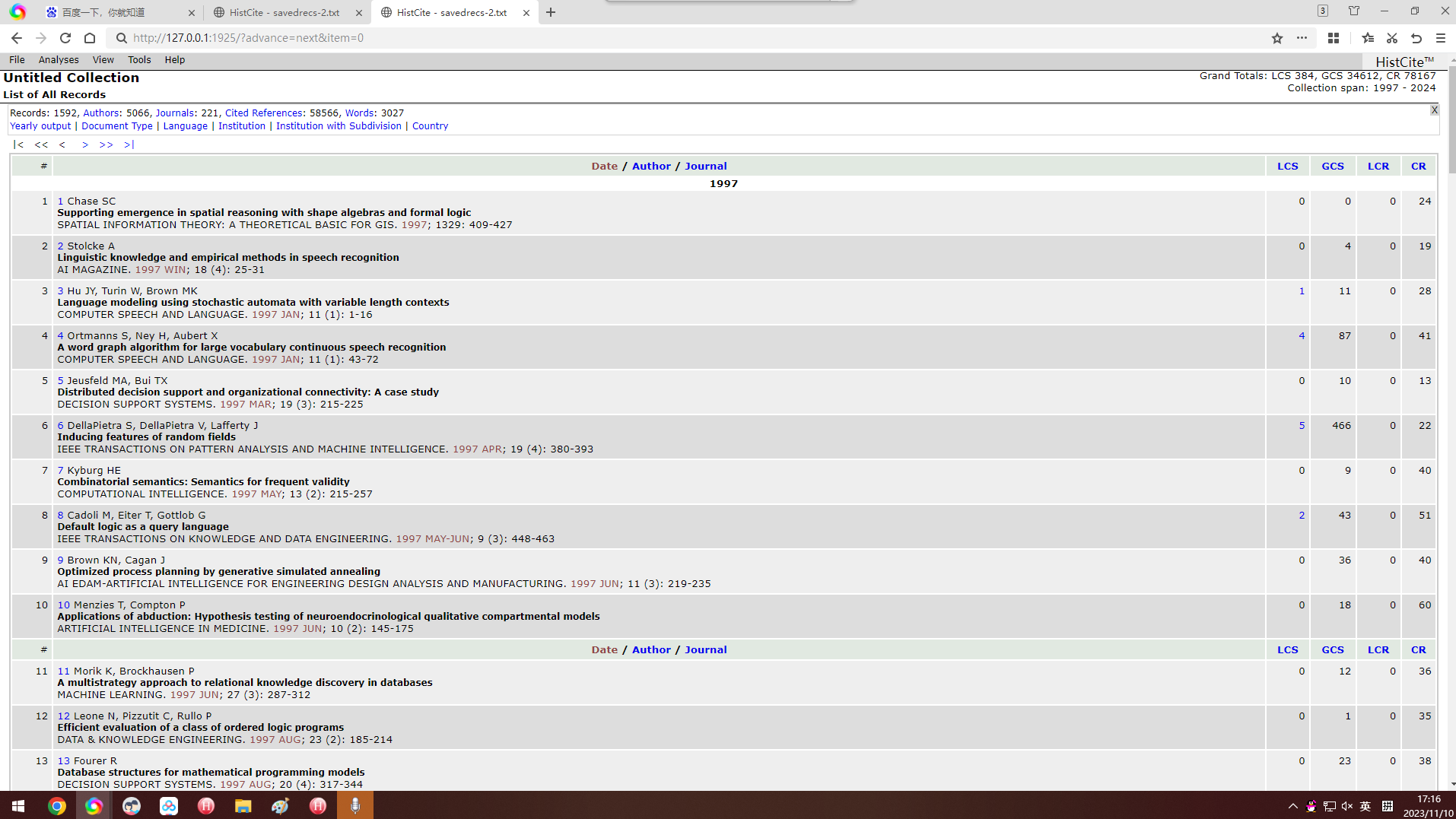


图3.4.1 导入成功

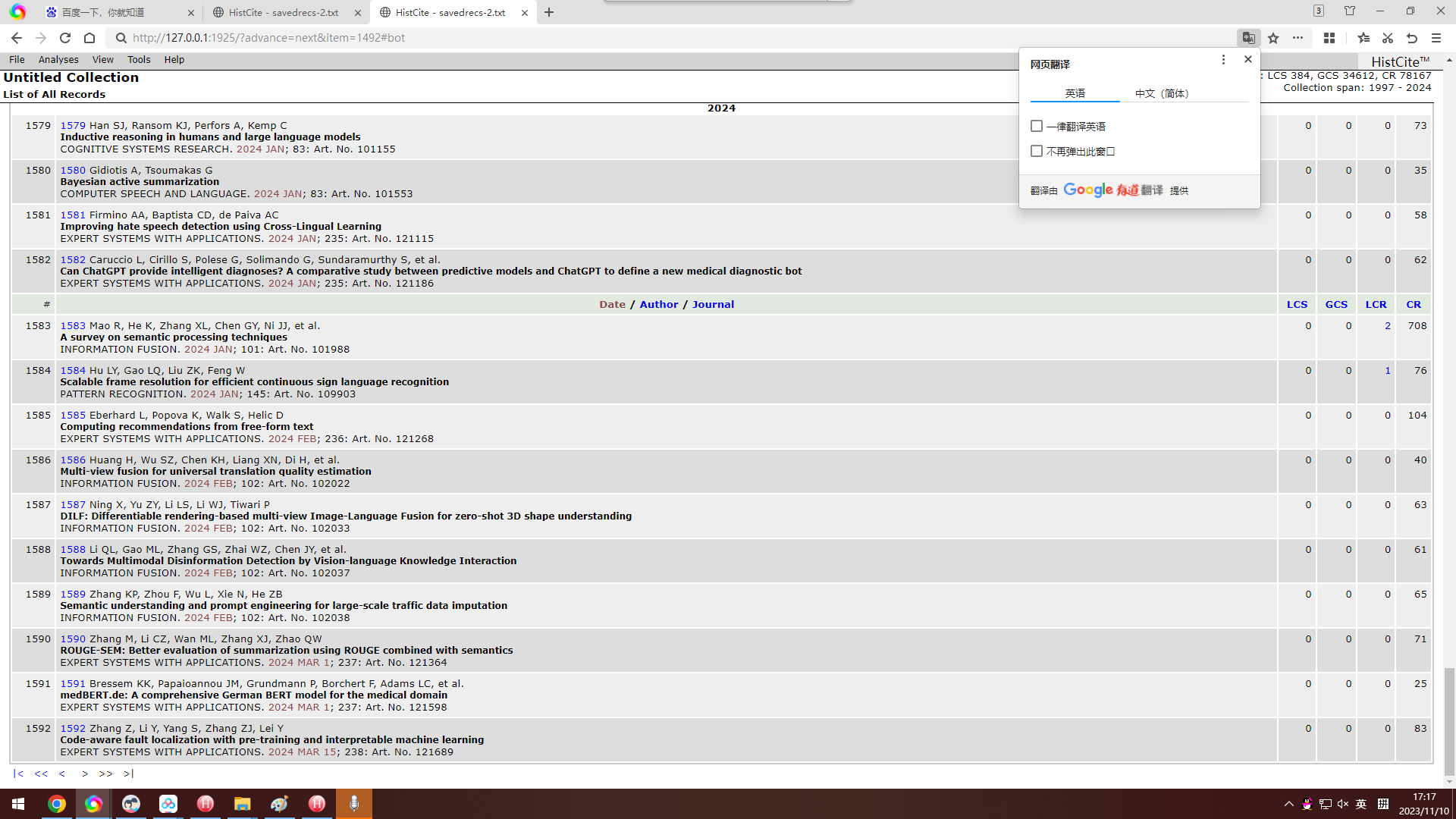


图3.4.2 导入成功

1. 作图分析：

Limit选择100，比较一下按 LCS 排序和按 gcs 排序的差异：

按 LCS 排序：

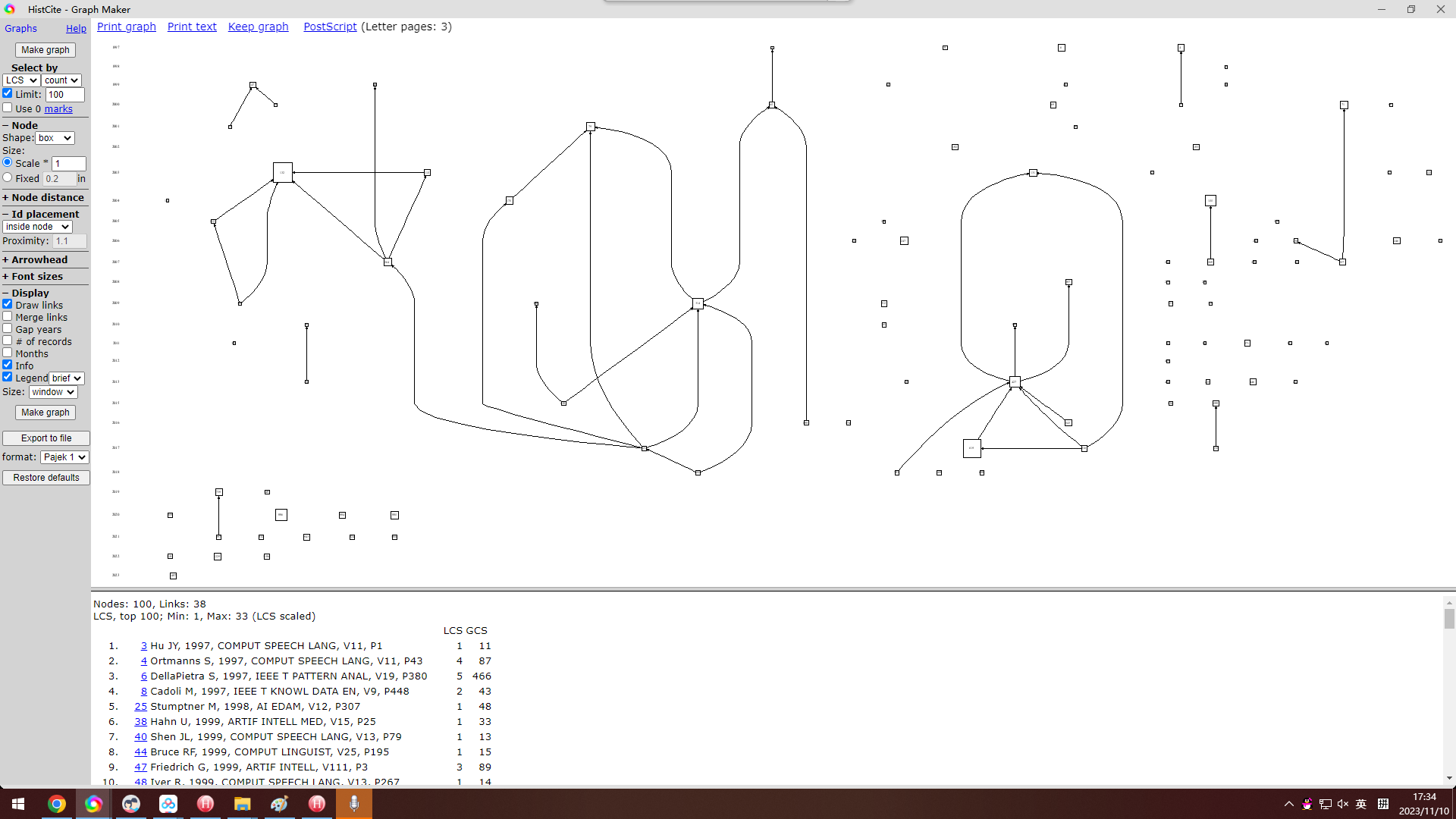


图3.5 Limit=100，按 LCS 排序

按 gcs 排序：

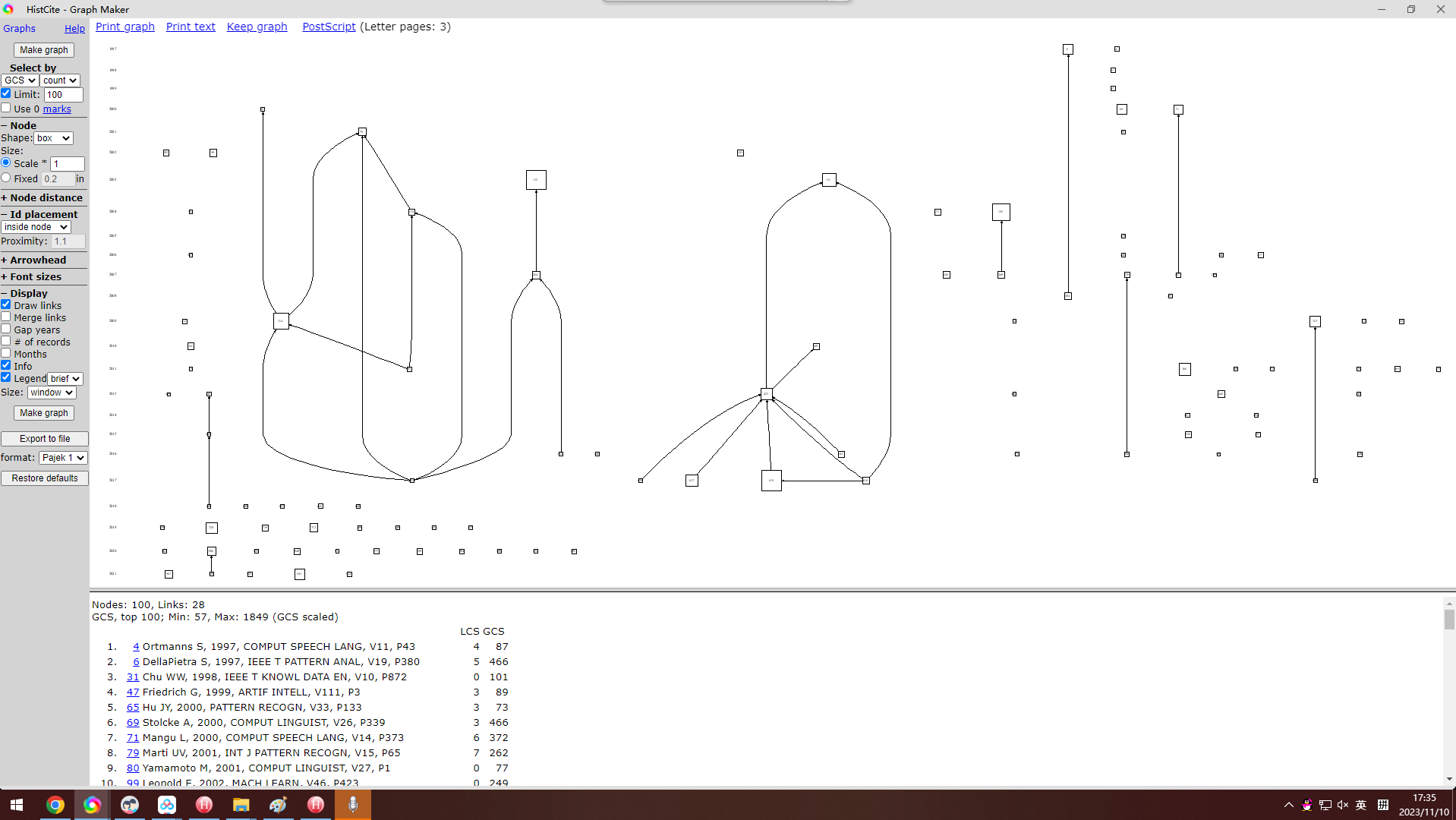


图3.6 Limit=100，按gcs排序

差异：LCS 作图，得到的关联较丰富，而 gcs 作图文献之间关联明显较少。

进行limit的选择，limit=200时关联性较强。

按 LCS 排序：



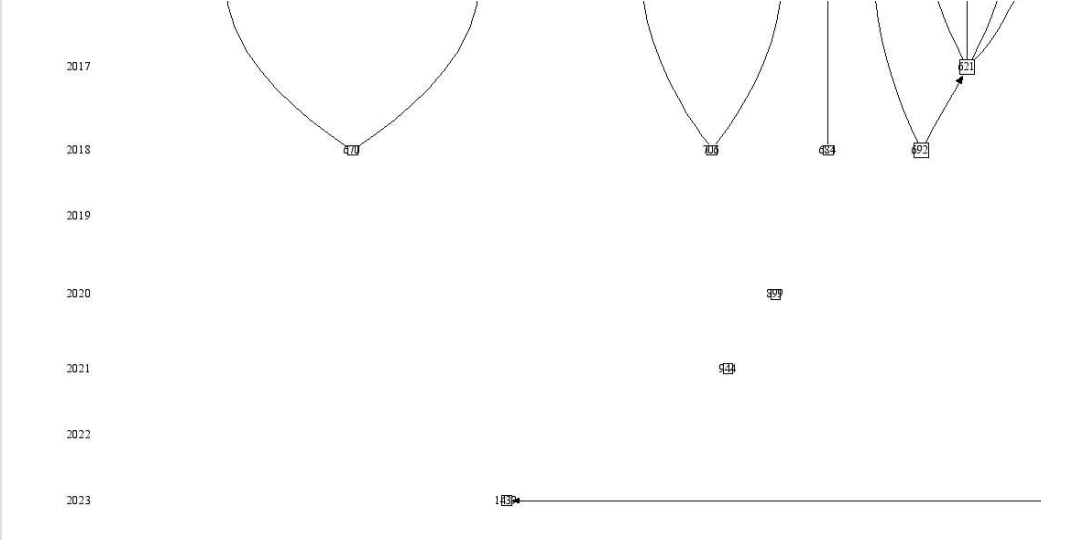


图3.7 limit=200，按 LCS 排序

分析：

2、3、4、5部分，不能只是贴截图，要把思路表达清楚，你要做什么，选了什么软件，相应功能的原理是什么，你怎么操作的，怎么调整的，这些都要说清楚。

关键步骤截图，所有图有编号有名称，有相应的解释文字。

在分析解释中，需要结合文献本身（可以略读文献标题、摘要）

## 4、多元统计分析

Spss部分

## 5、多视角的信息计量分析

（共词、合作，共被引、耦合，以及基于共现的聚类）

Ucinet/Gephi

Vosviewer

Citespace

这些基本的共现分析，都要求做到

## 6、结论

你这个研究话题的结论，需要结合前面的所有内容给出，也就是你用了一系列的信息计量方法，做了分析，最后得出的结论，应该给你在选题部分相呼应。2-3-4-5中的分析是结果分析，而6是结论。

## 7、实验心得

写你的经验、问题及解决、感想等等