小组成员：谢安琪、金成志、李鑫、陈思宇、王宏宇、任勇鹏、徐娆、周晖

摘要：

本次数据挖掘的实验数据是在学校图书馆的馆藏资源中通过Python爬虫进行爬取的，在爬取数据的同时也将每一篇论文的作者名称、期刊名称以及被引次数等相关信息一并爬取来方便后续进行分析。

我们小组这次打算要做的功能主要分为5个大模块，分别针对论文发表的年份、论文的被引次数、作者、高频词以及期刊进行分析。

针对论文发表的年份，主要是将每一年发表的论文个数统计出来，并且通过绘制一条折线图来直观反映出不同年份所发表的论文个数。从而能看出学术发展的整个趋势。

针对被引次数，主要是通过每一篇论文的被引次数，来对被引次数进行一个分段，0~5为一段，5~10为一段，以此类推，根据对不同被引次数的分段显示，用柱状图来进行显示，最终得到每个被引次数段有多少的数量，可以反映出这些数据集所能产生的价值有多大。

针对高频次数，主要是通过将这150篇论文进行分词，根据分出来的词就能进行统计并筛选出前50个高频词汇，最后通过词云图将这五十个单词显示出来。

针对作者信息，主要是通过每篇文章的不同作者，挖掘出不同作者之间有什么潜在的联系，并且在这150篇论文中，也存在同一批作者或者不同论文中的某几个作者一起合作之类的潜在问题，因此针对这一情况，所以挖掘出不同作者之间合作的信息。

这几种信息能够检测出论文的质量以及时效性情况，帮助用户更快地进行识别和参考适合自己的论文。最终可以将分析论文得到的结果进行下载和展示。

数据描述：

来源：成都信息工程大学图书馆馆藏资源，搜索关键词为“数据分析”，单位为成都信息工程大学

规模：150个数据集。