**数据工程与科学导论期末作业报告**

**10204304427 郝雨萌**

【实验名称】

酷狗音乐TOP500数据爬取处理可视化

【实验目的】

1、学习运用IPO过程处理数据。

2、学习爬虫的应用。

3、学习数据可视化的应用

【实验内容】

1. **需求分析**

（1）数据爬取：需要从酷狗音乐网站上爬取TOP500的音乐数据，包括歌曲名、歌手、专辑、排名等信息。

（2）数据处理：需要对爬取到的数据进行清洗、整理和格式化，去除重复数据、缺失数据等。

（3）数据可视化：需要将处理后的数据进行可视化，以便更好地展示音乐排名情况。可视化方式可以包括柱状图、折线图、词云图等。

**2、总体设计**

（1）数据爬取：使用 Python 中的 requests 库和 BeautifulSoup 库等工具来爬取酷狗音乐 TOP500 数据。通过分析酷狗音乐网站的网页结构和数据格式，定位到音乐排行榜数据所在的位置，并使用 requests 库发送网络请求获取数据。然后使用 BeautifulSoup 库解析网页内容，提取出需要的数据，如歌曲名、歌手、专辑、排名等信息。

（2）数据处理：对爬取到的数据进行清洗、整理和格式化，去除重复数据、缺失数据等。

（3）数据可视化：使用 Python 中的 Matplotlib、Wordcloud等库来进行数据可视化，如制作柱状图、词云图等。

**3、详细设计**

（1）数据爬取：

·使用 Python 的 requests 库发送 HTTP 请求获取酷狗音乐TOP500的网页源代码。

· 使用 BeautifulSoup 库解析网页源代码，提取出音乐排行榜数据所在的标签和属性。

· 遍历提取的标签和属性，获取每首歌曲的歌曲名、歌手、专辑和排名等信息，并保存到一个数据结构（如列表或字典）中。

（2）数据处理：

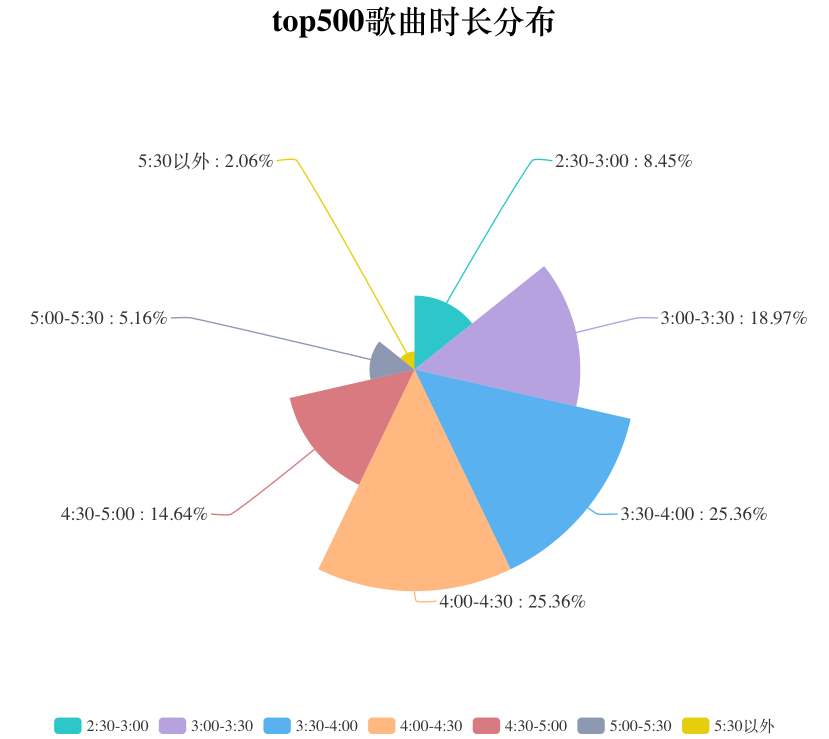
· 对爬取到的数据进行清洗和整理，用re正则提取出需要的歌词信息。

（3）数据可视化：

·使用 Matplotlib、Wordcloud库进行数据可视化。

·画出歌词词云图和歌手出现次数柱状图。

【数据可视化结果】

****