Python编程大作业方案及说明

方案概述： 本项目旨在设计一系列Python程序，解决一个实际问题：好事还是坏事更容易上热搜？中国有句古话“好事不出门，坏事传千里”。我们将基于定期爬取微博热榜获得的数据，运用人工智能分析对热榜新闻的情绪进行分析判断。我们首先定期爬取微博热榜获得数据，并将数据存储到一个数据库中。然后，对存储的数据运用人工智能和云计算进行处理和分析，并使用数据可视化技术生成图表。最后，设计一个网页，将生成的图表结果展示在网页上，以便用户能够方便地浏览和理解数据分析的结果。

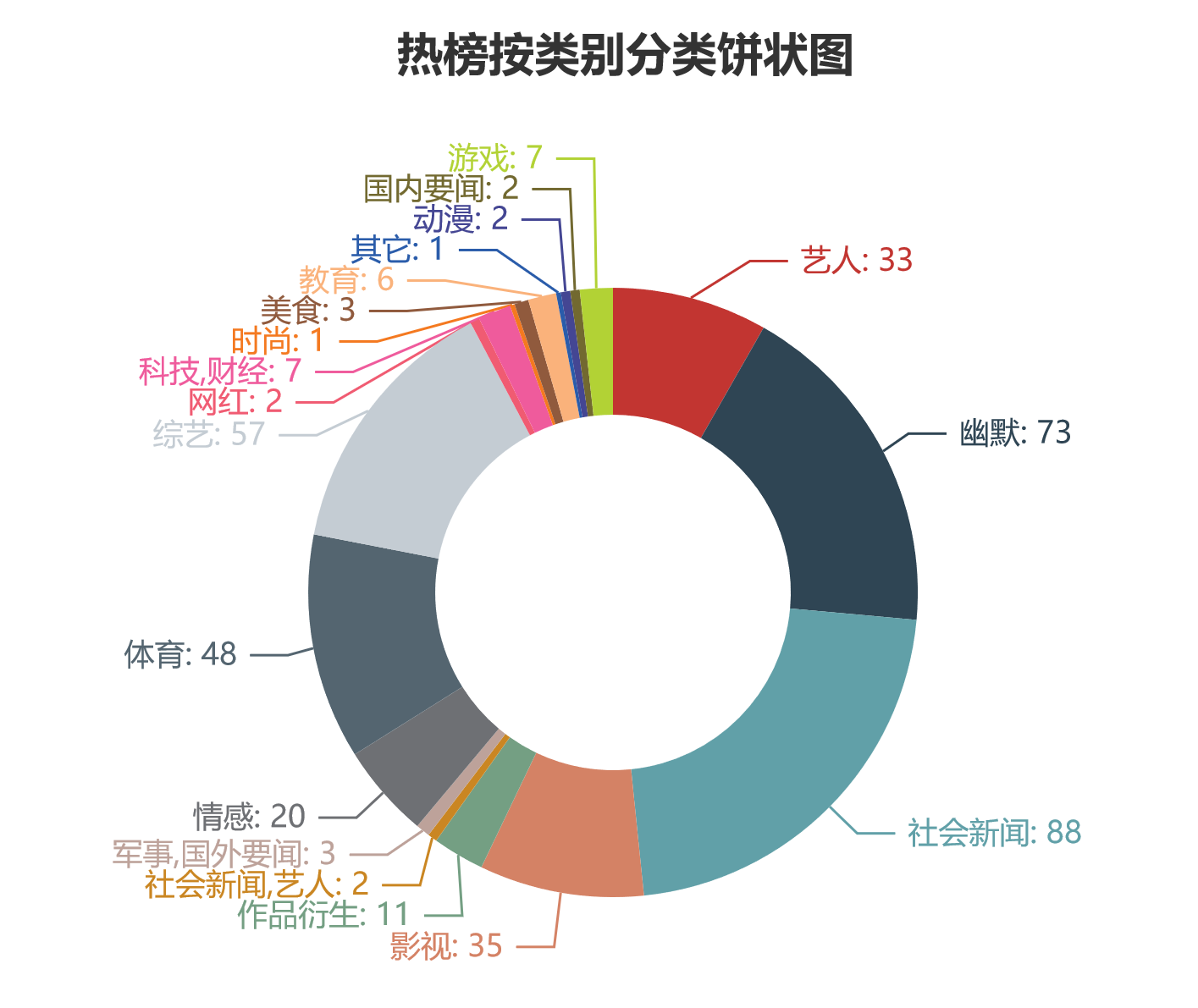
作业计划和说明：

1. 数据爬取：
   * 使用Python的爬虫库（如BeautifulSoup、Scrapy等）编写程序，实现对微博热榜页面的定期爬取。
   * 首先，需要了解微博热榜页面的结构和内容，确定所需信息的位置和获取方法。
   * 通过发送HTTP请求获取页面内容，并使用爬虫库解析HTML，提取所需的信息，如热榜标题、关键词、热度等。
   * 把写好的爬虫代码放到华为云上，让他跑两天，每隔一个小时（我们首先设定每分钟爬取一次作为尝试，每次大约获得20条数据，持续爬取二个小时144000条数据，但是由于发现微博热搜词的更新频率并不高，所以在之后收集的时候改为一个小时一次）爬取一次热搜榜
   * 将清洗后的数据存储到数据库中，可选择使用数据库管理系统（我使用了csv）进行存储。考虑使用pandas和numpy对数据进行处理。
2. 数据库设计：
   * 将爬取到的数据进行清洗和格式化，确保数据的准确性和一致性。
   * 根据爬取到的数据特点和需求，选择适合的数据库管理系统，并创建一个数据库用于存储微博热榜数据。
   * 设计数据库表结构，包括字段如热榜标题、关键词、热度、发布时间等。
   * 根据数据的关系和需求，设置适当的主键和索引。

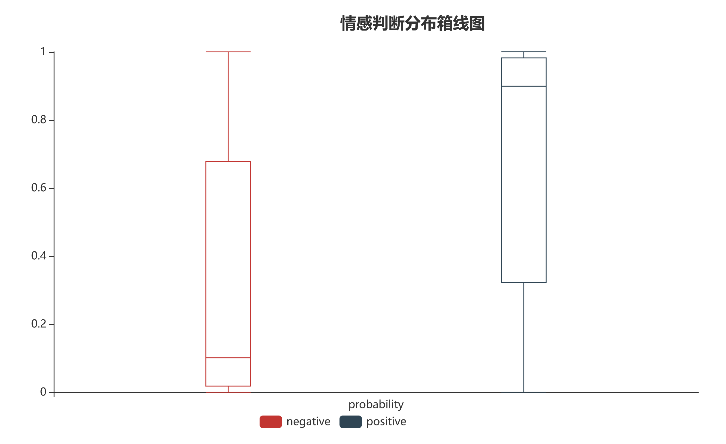
我主要将数据爬取下来之后整合成一个包含全部热搜信息的数据库，然后删除掉重复的热搜新闻。

1. 数据处理和分析（运用百度云）：
   * 使用Python编写代码，连接数据库，从数据库中提取爬取到的数据。
   * 针对数据特点和分析目标，进行数据处理和分析，以获得有价值的洞察。
   * 可进行数据清洗，如去除重复数据、处理缺失值等。
   * 根据需求进行数据聚合、排序、过滤等操作，以得出有关热榜的统计结果和趋势分析。
   * 百度云服务器给我提供了50万次的免费API接口，我获取了key和access token，然后将每一条的热搜新闻标题发给百度云的智能AI，然后返回一个结果，包含可信度（confidence）、判定为积极情感的概率（positive prob）、判定为消极情感的概率（negative prob）、判定的情感结果（sentiment，分为0、1、2三个结果，分别是消极的、中性的、积极的）
2. 数据可视化：
   * 使用Python的数据可视化库（如Matplotlib、Seaborn、Pyecharts等，主要使用pyecharts），将处理后的数据转化为图表形式。
   * 根据数据的类型和分析目标，选择适当的图表类型，如柱状图、折线图、饼图、热力图等。（考虑使用词云图对爬取的热搜进行描述性统计）
   * 通过图表展示数据的特征和趋势，使用户能够直观地理解和分析数据。
   * 我们主要做了四个图：

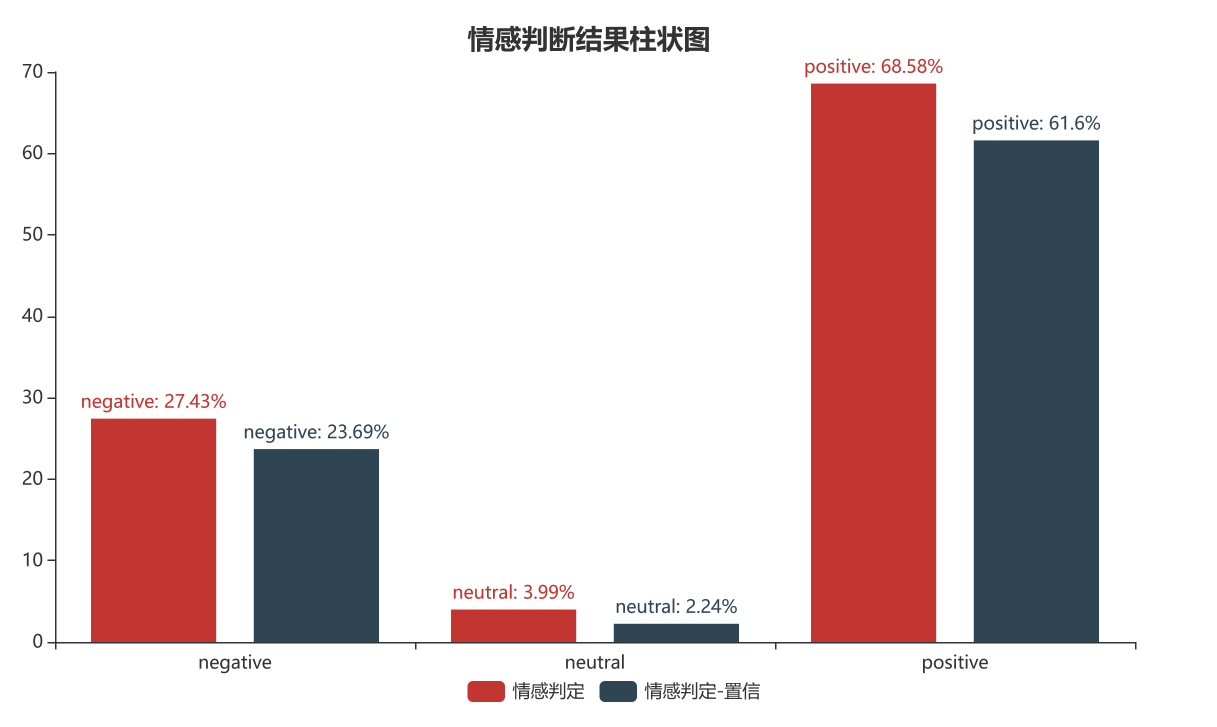
第一个图是对热搜类别占比的统计，可以看出，社会新闻、体育、幽默占比最大。



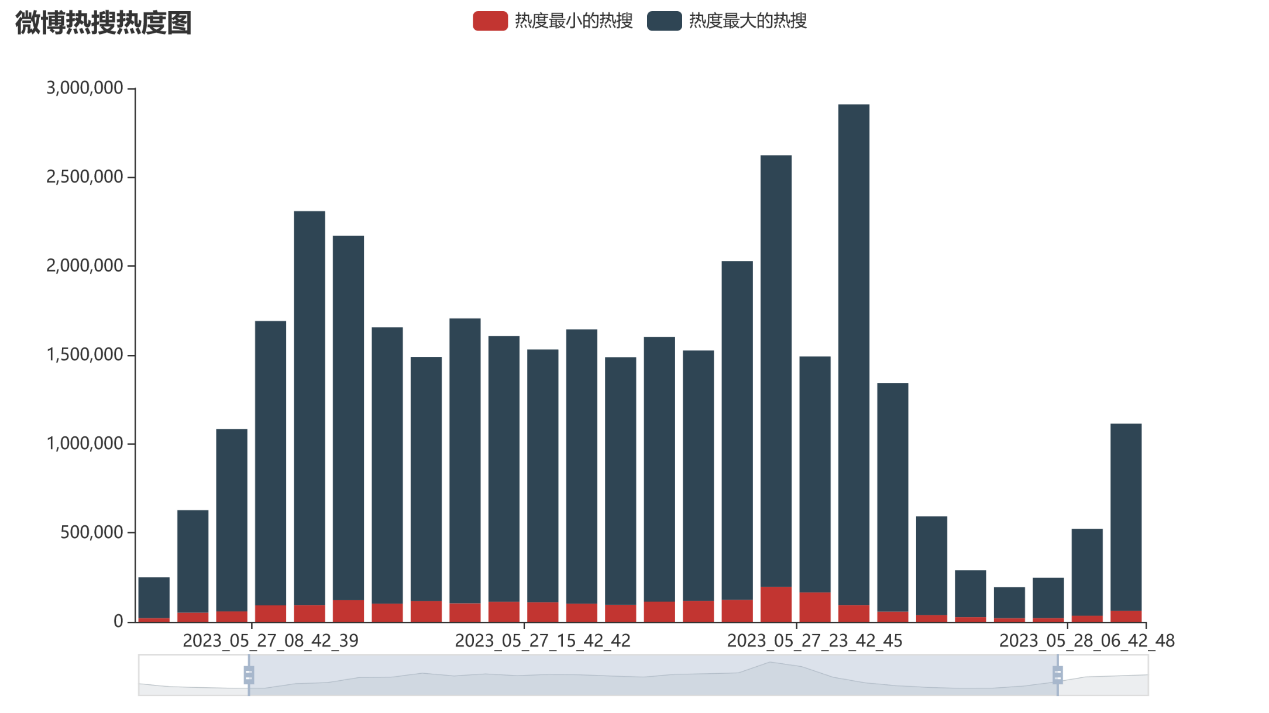
第二个图是情绪分析结果的分布图，用箱线图表示，可以发现，对于大部分热搜标题，判定为积极的可能性更大。



第三个图是对情绪判定结果的统计图，分为全部样本和置信程度在0.5以上的样本，可以看出，判定为积极的最多，消极的其次，中性的最少。



第四个图中我们对每隔一个小时爬取的热搜，对其热度最大和最小的分别进行统计，并增加了一个可以滑动的横坐标轴。



1. 网页设计与展示：
   * 使用HTML、CSS和JavaScript等前端技术，设计一个用户友好、美观大方的网。我用了一个现成的模板，修改了一下。