数据可视化大作业报告

小组课题：二次元数据可视化

姓名：勾王敏浩　　学号：2018214186

1.编程语言

python3

2.可视化任务

2.1 可视化任务 1

2.1.1 可视化目标

将每年的动漫数量做出总体统计，反映出动画产业在当年的增长变化。

2.1.2 数据来源

爬虫数据。

2.1.3 可视化思路

- 数据获取

  使用了 python 的爬虫库爬取了 D 站从 2010 年到 2018 年的动画数据。

- 可视化思路

  将爬取到的数据通过时间进行 group 操作，最终以散点图的形式展现。使用了散点图点的大小以及颜色作为视觉通道，对于 2010 年到 2018 年动漫数量的可视化结果如下：

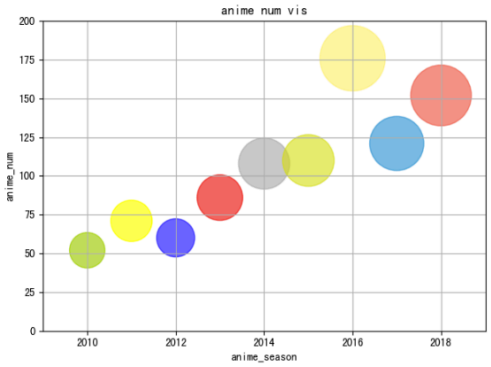

图 1 年度动漫数量可视化

2.1.4 简单分析

可以从图中直观观察到动漫数量随年份的变化，2010-2018 年终动画数量最终的三个年份分别为 2016，2018，2017.也能看到随着年份的增长，每年的动画数量有一个上升的趋势。

2.1.5 问题及解决

- 问题 1：在爬虫过程中，无法定位到特定需要选取的目标；
  解决 1：使用了 python 的 BeautifulSoup 以及正则表达式模块进行了精确地匹配。
- 问题 2：在绘制散点图过程中，报错颜色与 x，y 的 size 不匹配；
  解决 2：尝试使用了不同的方案，最终将 color 通过 operator 包进行扩展，使 size 相互进行匹配。

2.2 可视化任务 2

2.2.1 可视化目标

将每年的动画作品按照作品类型分类，反映出当年的动画产业在不同类型动画上的投资情况。

2.2.2 数据来源

爬虫数据。

2.2.3 可视化思路

- 数据获取

  使用了 python 的爬虫库爬取了 D 站从 2010 年到 2018 年的动画数据，包括了动画名称，动画播放时间以及动画标签。

- 数据分析

  对于爬取的动画类型维护了一个 set，检索了所有动漫数据所属的动漫类型。由于一个动画作品可能具备多个标签，将爬取到的第一个标签作为动画的主要标签。

- 可视化思路

  将需要查询年份的动画按照类型进行 group 最终进行统计，以饼图的形式展现了该年份不同动漫类型的占比。视觉通道主要是饼图扇形的大小以及不同区块的颜色。最终可视化效果如下：

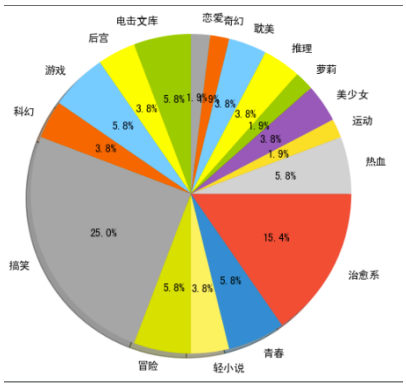

图 2 年度动画类型可视化

2.2.4 简单分析

从可视化结果可以直观看到不同年份不同类型动画所占据的类型，间接能够看出二次元爱好者的喜好变化。在本年度，数量最多的动画分别是搞笑类型以及治愈类型。

2.2.5 问题及解决

- 问题 1：显示中文乱码问题；

解决 1：修改了 python 中 plt 库的配置文件，并使用_rebuild 模块进行了重新加载。

2.3 可视化任务 3

2.3.1 可视化目标

可视化季度的动画热度。

2.3.2 数据来源

爬虫数据。

2.3.3 可视化思路

- 数据获取

  使用了 python 的爬虫库爬取了 D 站从 2010 年到 2018 年的动画热度信息其中热度以数值进行显示。

- 可视化思路

  将热度信息提取其中的数值进行可视化，为了更加明显的比较，采取了横向的直方图，以直方图的长度以及直方图的颜色作为视觉通道，最终可视化效果如下：

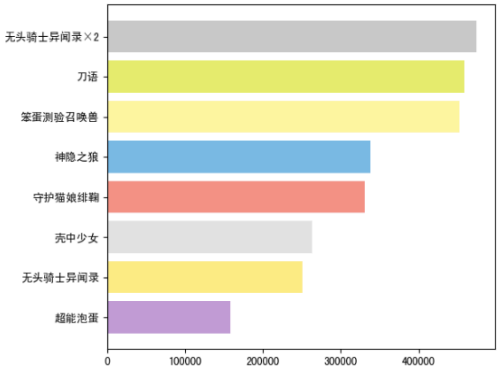

图 3 动画热度可视化

2.3.4 简单分析

从横向直方图中可以很直观地观察到动画的热度信息，对于动画的推荐、日后的制作方向能提供一定的指导。

2.3.5 问题及解决

- 问题 1：纵向的直方图对于排行信息的体现并不是很好；
  解决 1：采取了横向的直方图的方式，从上向下依序编码动画的热度。

2.4 可视化任务 4

2.4.1 可视化目标

使用词云将轻小说动画的部分章节可视化，提供动画的语义信息。

2.4.2 数据来源

《噬血狂袭》第三卷 《天使焚身》txt 文件

2.4.3 可视化思路

- 数据获取
  下载的动画轻小说，以 txt 文件进行保存。
- 可视化思路
  数据处理主要将下载的 txt 文件去掉空行。使用中文分词库 jieba 将 txt 分词，最终使用了 python 的 wordcloud 库形成了带有轮廓图的词云。最终可视化效果如下：

图 4 动画轻小说词云

2.4.4 简单分析

从词云中可以迅速把握到该部轻小说的主旨信息，是文本可视化一种非常好的工具。

2.4.5 问题及解决

- 问题 1：分词以及词云功能实现；
  解决 1：调研了很多库，最终选用了 python 的 jieba 以及 wordcloud 进行了实现。

2.5 可视化任务 5

2.5.1 可视化目标

可视化动画轻小说中的人物关系。

2.5.2 数据来源

《噬血狂袭》第三卷 《天使焚身》txt 文件

2.5.3 可视化思路

- 数据获取

  下载的动画轻小说，以 txt 文件进行保存。

- 可视化思路

  将轻小说的人名独立为字典，使用中文分词库 jieba 将文本进行分词，找出其中的人名，如果两个人名出现在相同的段落，就认为这两个角色之间具有关联性，最终将人物的共现关系使用社交网络库 networkx 进行可视化，可视化通道为边的颜色（共现的次数），对于不同的角色联系使用了不同颜色的边进行了标识。可视化效果如下：

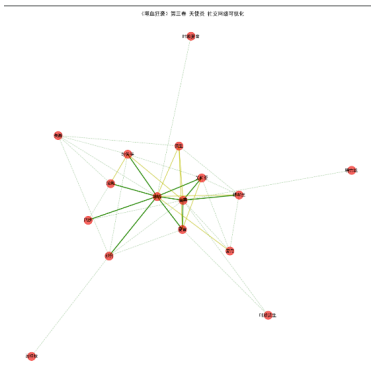

图 5 动画社交网络关系可视化

2.5.4 简单分析

从社交网络图中可以清晰的捕捉到人物的重要程度以及人物之间的联系。

2.5.5 问题及解决

- 问题 1：设计方面如何捕捉到两个人物之间的关联程度；

  解决 1：使用了人物的共现关系，如果两个人物在同一段落共同出现的次数越多，就认为这两个人物之间的相关程度越高。

- 问题 2：指定日文人物名字；

  解决 2：由于日文的名字与中文的名字差别较大，使用 jieba 分词可能出现分词不准确的情况，所以对于本章节可能出现的所有人名，维护了一个 jieba 分词库能够识别的词典。