

摘要

仰望星空，漫天的星斗，它们尽着自己的力量，把点点滴滴的光芒融汇在一起，虽然比不上太阳的辉煌，也比不上月亮的清澈，但他们梦幻般的光，洒到了人间，照亮了人们的心头。我一边走一边看，发现了天蝎、射手、水瓶.....12 星座在夜晚的闪耀为星象学家预测运势打开了光明的大门。在科学技术迅速发展的今天，星座分析与数据科学又能够碰撞出什么火花呢？

关键词：12 星座；弗雷歇距离；拟合优度；数据可视化

Abstract

Looking up at the starry sky and the stars in the sky, they do their best to blend the light bit by bit. Although they are not as brilliant as the sun or as clear as the moon, their dreamlike light shines on the world and illuminates people's hearts. While walking, I found Scorpio, Sagittarius, Aquarius The shining of the constellations at night opens a bright door for astrologers to predict their fortunes. With the rapid development of science and technology, what sparks can constellation analysis and data science collide with each other?

Keywords: 12 constellations; Frechet Distance; Goodness of Fit; Data visualization

目录

摘要	1
1. 引言	3
1.1 项目背景	3
1.1.1 选题背景	3
1.1.2 研究背景	3
1.2 问题描述	3
1.3 研究方法 & 难点	4
1.3.1 研究方法	4
1.3.2 研究难点	4
2. 实验工作与结果分析	5
2.1 数据爬取与处理	5
2.1.1 星座屋网站的数据爬取	5
2.1.2 闹闹每日星运的数据爬取	6
2.1.3 星座女巫 Tarot 的数据爬取	8
2.1.4 针对爱情运势文本词云展示的数据处理	9
2.1.5 针对契合星座判断、幸运颜色拟合、爱情运势得分变化模拟的数据处理	10
2.1.6 基于 Frechet Distance 的时段运势轨迹相似性模型和基于拟合优度的运势打分拟合模型的数据处理	10
2.2 各星座爱情运势分析与可视化	12
2.2.1 爱情运势文本词云展示	12
2.2.2 基于动态柱状图的契合星座判断	12
2.2.3 基于每日爱情运势得分的图象模拟	14
2.3 基于饼状图的幸运颜色拟合	15
2.4 基于 Frechet Distance 的时段运势轨迹相似性模型	16
2.5 基于幸运数字的拟合优度模型	19
3. 结论与展望	22
4. 前期未成形模型	22
5. 致谢	24
6. 引用	24

1.引言

1.1 项目背景

1.1.1 选题背景

“你会为了我去搜索陶白白吗？”这一句话，让抖音上的星座情感博主@陶白白 Sensei 同时在抖音和微博登上了热搜榜。在那之后，陶白白越来越出圈，一时间成了全网、多平台增粉最快的博主，目前在微博与抖音平台，都拥有千万级粉丝。伴随着陶白白的爆火，“都 1202 年了，怎么星座博主又火了？”这一趣味性热梗也在各大娱乐平台引发了关于星座话题的广泛讨论。

百度指数显示，在星座、塔罗、周易等耳熟能详的占卜方法中，星座的热度就像 2018 年蔡徐坤的出道情景——断崖式领先。从近 5 年的平均搜索热度来看，星座不仅排名第 1，而且是第 2 名周易的 13.5 倍。小编我本人也是星座圈的狂热分子，陶白白的爆火激起了我用数据化、科学化方式研究这一领域的热情。

1.1.2 研究背景

星座学最早诞生于公元 3 千年，其中心思想是认为天体的位置和运动会直接影响人的性格、生活、甚至命运等。星象学的发展和人类早期对于天文学的探索有着千丝万缕的联系。但是，随着科学对于实证越来越高的要求，星象学与天文学日渐区分开来。现在的星座学与天文学互不相干。因而为了保险，星座圈还流行着一句话，“如果你信星座，那你就不能只看陶白白”，意思是说，我们需要多看几种星座分析，综合比对。

从哲学的角度来讲，星象运动是固定的，存在决定意识，不同博主的星象分析也应该是一致的，因为个人偏好可能会略有偏差，但绝不应该是完全不同的。如果通过研究我们发现各位博主的分析基本趋于一致，也更加能验证星座运势是可以相信的。那么来自不同出处的星象分析是否趋于一致呢？我将通过模型建立、数据可视化等方式来深入探索研究。

1.2 问题描述

首先基于星座圈特别火的一句话“看星座的人，80%看爱情”，我将从各个维度综合展示从 2022 年 9 月 1 日到 2022 年 12 月 20 日的爱情运势情况。

之后也是本次实验最重要的内容，我将基于不同博主对同一运势不同维度的分析建立不同的数学模型，进行比对和拟合。

1.3 研究方法及难点

1.3.1 研究方法



图 1 实验研究流程

- ① 在 Windows 环境下，利用 vscode 和 Anaconda 来爬取一个网站和两个微信公众号的数据。
- ② 调用 pandas、matplotlib、numpy、jieba、wordcloud 等库来可视化数据和建立模型。

1.3.2 研究难点

在数据爬取方面，选取的网站为星座屋，微信公众号为闹闹每日星运、星座女巫 Tarot。星座本就是一个很小众的话题，计算机科学等学科在该领域的研究更是少之又少，所有的爬虫代码需自己编写，而数据来源又来自三个不同的平台，需要编写三类结构不同的爬虫代码。星座女巫 Tarot 的反爬虫技术及其强大，几乎每 7 天就会换一个版本，因而只是为了爬取到星座女巫上充足的数据，就写了 7 个版本的爬虫代码，再加上其他两个数据来源，一共写了 10 篇爬虫的代码。

在数据可视化方面，由于爬虫技术在之前并没有接触过，项目期间短期的学习并不十分熟练，爬下来的数据过于繁多和冗杂，因而需要大量的工作进行数据分类和清洗。数据类型有文本，有数字，需要不同的可视化方式。

在模型建立方面，本次实验，一共建立了两个模型：基于 Frechet Distance 的时段运势轨迹相似性模型和基于幸运数字的拟合优度模型。第一个模型，在知网和谷歌学术上找到的有关 Frechet Distance 的算法研究很少，在翻阅了十几篇文章后，找到了一篇关于 Frechet Distance 算法的近似拟合的伪代码，因而独立编写复现了这篇代码。第二个模型是在查阅了大量的数学文献资料后选定了针对特殊数据最为合适的评判标准，找出相关数学公式建立模型。

2.实验工作与结果分析

2.1 数据爬取与处理

2.1.1 星座屋网站的数据爬取

之所以选取星座屋，是因为当下虽然随着智能手机的广泛普及，网站的浏览量有所下降。但是该网站是星座圈人士认为内容最丰富、分析最准确的网站，也是在曾经以电脑为主要上网媒介的年代单日浏览量最高的星象网站。

该网站链接为 <https://www.xzw.com/fortune/>，进入以后我们先分析网站的结构。该网站包含运势、生肖、塔罗牌等多种内容资源，而我需要的是对 12 星座每日运势的分析，通过鼠标右键检查找到日运势所在的模块。



图 2 星座屋网站

之后我通过 headers 等信息处理网站的反爬机制，然后通过循环对每日运势不同方面的分析逐个爬取，存储到 csv 文件中。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
标题	综合运势：爱情运势：事业学业：财富运势：健康指数：商谈指数：幸运颜色：幸运数字：速配星座：短评：															综合运势：爱情运势：事业学业：财富运势：健康运势文字：
153 狮子座明晨	48	48	48	32	71%	58%	卡其色	5 摩羯座	凡事掌握一低迷的运势单身的女性尽量在小溪没有理智的时候务必							
154 狮子座明晨	64	48	64	64	82%	76%	墨绿色	8 狮子座	迎来一些稍稍有起伏的单身或者需要进行正财或能建议保证充足的							
155 狮子座明晨	64	64	32	64	80%	81%	米黄色	1 天蝎座	再接再厉的运势尚可，单身的主动表面看起来会主财一些消化能力有所下							
156 狮子座明晨	64	64	64	64	92%	94%	宝蓝色	7 水瓶座	以远瞻目光运势佳单身或者能够寻求财的财求喝水的时候不要							
157 狮子座明晨	48	48	64	48	80%	72%	浅棕色	6 金牛座	时间的积累运势一般，单身的容易通过努力求财财运平平，日常生活中尽量							
158 狮子座明晨	64	64	64	64	96%	92%	紫色	3 天秤座	出击势头运势特别佳单身的单身总是保持着财运正旺，建议保持早睡早							
159 狮子座明晨	48	48	32	64	68%	61%	杏色	2 白羊座	表现愈显佳整体运势佳单身的单身需要在晚上没有明显建议做好身体方							
160 狮子座明晨	48	48	48	32	72%	74%	墨绿色	5 水瓶座	迫切渴望运势稍有单身的单身需要从不同投机取巧进行情绪方面的							
161 狮子座明晨	48	32	48	32	69%	55%	玫红色	7 天蝎座	得过且过的运势较差，单身的切勿容易出解解易于求财由于睡眠不足的							
162 狮子座明晨	64	64	48	64	88%	80%	蓝色	8 双鱼座	坚定前行的运势还不单身的有逐步优越走步容易得到小避免用脑过度的							
163 狮子座明晨	64	64	64	64	93%	90%	米白色	1 狮子座	闯出自己的整体运势佳单身的单身拥有一定的财运不错，走路的时候需要							
164 狮子座明晨	64	64	64	48	82%	88%	浅绿色	4 处女座	创造更多的运势尚可，单身的单身需要忙碌的财运平平，或有养生的想法							
165 狮子座明晨	48	48	48	48	73%	62%	青色	9 射手座	重视关系的整体运势佳单身的单身需要考验人求财相对生活中磕磕撞撞							
166 狮子座明晨	48	48	64	48	79%	85%	焦糖色	2 金牛座	专注于心灵运势稍有单身的单身一心一用财运一般能改掉生活中的不							
167 狮子座明晨	32	32	48	32	71%	57%	棕色	3 天蝎座	抠细节的运势低迷，单身的单身对自己要求求财多需要才需要通过正确的							
168 狮子座明晨	48	48	32	48	69%	54%	朱红色	7 水瓶座	大脑处于迷运势欠佳，单身的单身需要慢慢求财财阻力建议保持均衡以							
169 狮子座明晨	48	64	48	64	84%	74%	粉色	6 双子座	主动迎接运势尚可，单身的单身警惕从众才能够在稳长时间处于噪音							

图 3 星座屋爬取的数据

```

for u in url_list:
    response = requests.get(url=u, headers=headers)
    print(response)
    soupi = BeautifulSoup(response.text, 'lxml')
    # 解析页面
    try:
        dic = {}
        dic['标题'] = soupi.find('h4').text
        # 获取标题信息
        infor1 = soupi.find('div', class_='c_main').find('ul').find_all('li')
        dic[infor1[0].text] = infor1[0].find('em')['style'].split(':')[1].split('p')[0]
        dic[infor1[1].text] = infor1[1].find('em')['style'].split(':')[1].split('p')[0]
        dic[infor1[2].text] = infor1[2].find('em')['style'].split(':')[1].split('p')[0]
        dic[infor1[3].text] = infor1[3].find('em')['style'].split(':')[1].split('p')[0]
        dic[infor1[4].find('label').text] = infor1[4].text.split(':')[1]
        dic[infor1[5].find('label').text] = infor1[5].text.split(':')[1]
        dic[infor1[6].find('label').text] = infor1[6].text.split(':')[1]
        dic[infor1[7].find('label').text] = infor1[7].text.split(':')[1]
        dic[infor1[8].find('label').text] = infor1[8].text.split(':')[1]
        dic[infor1[9].find('label').text] = infor1[9].text.split(':')[1]
        # 获取运势等信息
        infor2 = soupi.find('div', class_='c_cont').find_all('p')
        dic['综合运势文字'] = infor2[0].find('span').text
        dic['爱情运势文字'] = infor2[1].find('span').text
        dic['事业学业文字'] = infor2[2].find('span').text
        dic['财富运势文字'] = infor2[3].find('span').text
        dic['健康运势文字'] = infor2[4].find('span').text
        print(dic)
        data_list.append(dic)

```

图 4 星座屋核心爬虫代码

2.1.2 闹闹每日星运的数据爬取

闹闹每日星运主打的就是丰富精准的每日运势、每周运势、每月运势，每日运势文章当天浏览量都能超过 3 万日次，收获粉丝好评。



图 5 闹闹每日星运网站

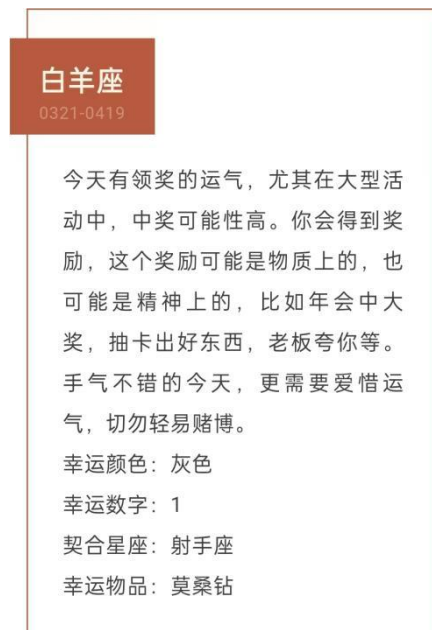


图 6 闹闹每日星运日运势所含内容

微信公众号文章的爬取与网站有着截然不同的形式。网站传播本就靠链接的转载，而微信公众号文章的传播是靠转发，我们在正常情况下无法获得直接指向该公众号的链接。为此，我们要钻一个微信公众号文章制作图文时引用其他文章的一个小空子。原理就是，当我们登录微信公众号后台，进行图文素材编辑的时候，可以在素材中插入其他公众号的推送链接。这里微信公众号后台会自动调用相关 API，返回该公众号所有推送的长期链接列表。我们打开 Chrome 浏览器的检查模式，选择 Network，然后在编辑超链接界面的公众号搜索栏中输入“闹闹每日星运”，搜索并选择该公众号，发现 Network 中刷新出了一个开头为“appmsg”开头的内容，这就是我们要分析的目标。点击“appmsg”开头的这条内容，解析该请求的 url，获取 cookies、User-Agent 等参数，存入一个 yml 文件，然后获取该公众号从 2022 年 9 月 1 日到 2022 年 12 月 20 日每日运势的所有链接。

```
while True:
    begin = i * 5
    params["begin"] = str(begin)
    # 随机暂停几秒，避免过快的请求导致过快的被查到
    time.sleep(random.randint(1,10))
    requests.packages.urllib3.disable_warnings()
    resp = requests.get(url, headers=headers, params = params, verify=False)
    # 微信流量控制，退出
    if resp.json()['base_resp']['ret'] == 200013:
        print("frequency control, stop at {}".format(str(begin)))
        time.sleep(3600)
        continue

    # 如果返回的内容中为空则结束
    if len(resp.json()['app_msg_list']) == 0:
        print("all article parsed")
        break

    msg = resp.json()
    if "app_msg_list" in msg:
        for item in msg["app_msg_list"]:
            info = "{}", "{}", "{}", "{}".format(str(item["aid"]), item['title'], item['link'], str(item['create_time']))
            with open("app_msg_list_naonao.csv", "a", encoding='utf-8') as f:
                f.write(info+'\n')
            print(f"第{i}页爬取成功\n")
            print("\n".join(info.split(",")))
            print("\n\n-----\n")

    # 翻页
    i += 1
```

图 7 闹闹每日星运日运势链接批量爬取代码

当爬取 50 页左右时，遇到如下错误：{'base_resp': {'err_msg': 'freq control', 'ret': 200013}} 这是因为微信公众号存在流量限制，等待一小时即可。我这里采用如下代码解决。

```
# 微信流量控制，退出
if resp.json()['base_resp']['ret'] == 200013:
    print("frequency control, stop at {}".format(str(begin)))
    time.sleep(3600)
    continue
```

图 8 微信流量控制应对代码

,aid,title,url,time	
0,2658509243_2,星历1220:	巨蟹适合积累知识 狮子顶住压力焦虑,http://mp.weixin.qq.c
1,2658509205_2,星历1219:	水瓶排除错误信息 金牛注意呼吸系统,http://mp.weixin.qq.c
2,2658509190_2,星历1218:	巨蟹避免头疼 双鱼开始养生,http://mp.weixin.qq.com/s?_b
3,2658509155_2,星历1217:	天蝎适合参加聚会 处女要拒绝冷战,http://mp.weixin.qq.com
4,2658509143_2,星历1216:	天蝎不要委屈身体 金牛挖掘自己潜力,http://mp.weixin.qq.c
5,2658509112_2,星历1215:	摩羯保持乐观 双鱼开始养生,http://mp.weixin.qq.com/s?_b
6,2658509094_2,星历1214:	天秤保持平和心态 水瓶建议好好休息,http://mp.weixin.qq.c
7,2658509062_2,星历1213:	射手主动改变 金牛外出交友,http://mp.weixin.qq.com/s?_b
8,2658509051_2,星历1212:	水瓶财运好转 巨蟹适时沉默,http://mp.weixin.qq.com/s?_b
9,2658509009_2,星历1211:	摩羯加强日常锻炼 狮子避免旧事重提,http://mp.weixin.qq.c
10,2658509005_2,星历1210:	处女参加慈善事业 双鱼保护自己财物,http://mp.weixin.qq.c
11,2658508995_2,星历1209:	巨蟹需要断舍离 双子拓展个人事业,ht 打开链接 (ctrl + 单击)
12,2658508993_2,星历1208:	摩羯谨慎投资理财 射手平衡各方关系,http://mp.weixin.qq.c
13,2658508958_2,星历1207:	双子注意说话尺度 天秤注意心理健康,http://mp.weixin.qq.c
14,2658508900_2,星历1205:	水瓶注意安全 白羊保持低调,http://mp.weixin.qq.com/s?

图 9 闹闹每日星运爬取到的日运势链接

之后，我们从 csv 文件中读取每篇推送的 url 链接，用 Requests 库爬取每篇推送的内容。

```
if i>0:
    response = requests.get(url,headers=headers)
    soupi = BeautifulSoup(response.text, 'html5lib')
    try:
        # 获取标题信息
        a = soupi.find('h1').text
        #print(a)
        a1 = a.strip()
        a1 = a1[2:6]
        #print(a1)
        infor01 = soupi.find('div', class_="rich_media").find('div',class_="rich_media_inner").find('div',class_="rich_media_area_primary").find
        #print(infor01)
        infor1 = infor01.find("section").next_sibling.next_sibling.next_sibling
        for j in range(1,13):
            dic = {}
            a2 = infor1.find('section').find('section').find('section')
            a3 = a2.find("section").find('section').find("section").find("section").find("p").find('strong').text
            dic['标题'] = a3 + a1 + "运势"
            a4 = a2.find('section').next_sibling.find("section")
            temp = a4.find("section")
            dic['综合运势'] = temp.find('span').text
            temp = temp.next_sibling
            dic['幸运颜色'] = temp.find('span').text
            temp = temp.next_sibling
            dic['幸运数字'] = temp.find('span').text
            temp = temp.next_sibling
            dic['契合星座'] = temp.find('span').text
            temp = temp.next_sibling
            dic['幸运物品'] = temp.find('span').text
            print(dic)
            data_list.append(dic)
            #print(data_list)
            #print("/n")
            #number = number + 1
            infor1 = infor1.next_sibling
```

图 10 闹闹每日星运日运势内容爬取核心代码

2.1.3 星座女巫 Tarot 的数据爬取

星座女巫 Tarot 主打的是对每日运势从爱情、事业学业、健康三个不同的维度分析，虽然在星座圈里这个公众号很小众，但是据一部分看过的网友分析，该公众号是具有一定可信度的。

该公众号的反爬机制是三个数据源中做的最好的，在 9 月和 10 月差不多每

隔 7 天就会换一种反爬机制，针对这个原因，我不得不编写 7 个不同的爬虫版本来爬取到足够的数据。直到今天，我一想到那一个星期每天改爬虫代码改到吐的场景都会留下一把辛酸泪。由于星座女巫 Tarot 与闹闹每日星运同属于公众号，爬取的大方向是一致的，我将不再赘述。



图 11 星座女巫 Tarot 每日运势内容

2.1.4 针对爱情运势文本词云展示的数据处理

首先根据标题截取出 12 个星座的名称，然后挑选出某个星座某一特定时期的爱情运势文本，运用 python 的 join 函数对文本进行切片，之后调用 jieba 分词对不同的切片进行具体的分词处理。

```
def ciyun():
    data["标题"] = data["标题"].str[:3] # 首先要将标题只截取出来星座的名字
    all_fortune = data[data["标题"] == "天秤座"]["爱情运势文字"].tolist()
    print(all_fortune)
    all_fortune_txt = "".join(all_fortune)
    #print(all_fortune)
    #将综合运势的文字进行切片
    word_list = jieba.cut(all_fortune_txt)
    txt = "".join(word_list)
    print(txt)
```

图 12 词云图所需数据的清洗代码

2.1.5 针对契合星座判断、幸运颜色拟合、爱情运势得分变化模拟的数据处理

虽然契合星座判断、幸运颜色拟合、爱情运势得分变化模拟属于本次实验三个不同的内容，但是三者数据清洗的方式是一致的，因此我将以幸运颜色拟合为例详细讲解。

通过 pandas 的 read_csv 函数打开对应的数据文件，由于标题内容蕴含丰富，我们首先对标题切片，分别一一对应储存好星座名称和日期，然后根据传入的参数 name（幸运颜色/幸运星座/爱情运势）找到相应的数据栏，记录下该时段某个星座或某种颜色的出现频率，以字典对的形式加入列表。值得注意的是，由于在数据爬取时并没有注意到 9 月和 10-12 月数据存取是以不同方式，因而在数据清洗时，需要写两种代码。

```
def datawash1(content_xingzuowu, Name, num, name):
    #星座屋
    data = pd.read_csv('xingzuo.csv', encoding='utf-8', )
    data_1 = data.set_index("标题", drop=False)
    data_1.index = data_1.index.str[7:]
    #print(data_1.index)
    data_1.index.name = "date"
    data_1["标题"] = data_1["标题"].str[:3] #水瓶座
    #print(data_1["标题"])
    data_2 = data_1[data_1["标题"] == Name]
    #print(data_2)
    data_3 = data_2[content_xingzuowu].tolist()
    data_3 = Counter(data_3)
    n = len(data_3)
    name = list(data_3)
    #print(data_3)
    #print(name)
    print(n)
    for i in range(0, n):
        n = name[i]
        #print(data_3[i])
        num.append(int(data_3[n]))
    #print(data_3)
    #print(name)
    #print(num)
    xingzuowu = []
    xingzuowu.append(name)
    xingzuowu.append(num)
    return xingzuowu
```

图 13 所需的 9 月数据清洗核心代码

2.1.6 基于 Frechet Distance 的时段运势轨迹相似性模型和基于幸运数字的拟合优度模型的数据处理

基于 Frechet Distance 的时段运势轨迹相似性模型模型选取的数据来源是星座女巫 Tarot 和星座屋网站，基于拟合优度的运势打分拟合模型选取的数据源是星座屋网站和闹闹每日星运微信公众号。虽然两个模型所需数据来源，数据内容不同，但是数据清洗方式相似，我将以基于 Frechet Distance 的时段运势轨迹相

似性模型为例进行分析。由于数据源的格式不同，需要编写两个不同的函数清洗数据，但是数据清洗的大方向是一致的。因此我将以星座女巫 Tarot 为例进行详细介绍。前面提到 9 月和 10-12 月两个时段的数据存取方式也不同，我将以 9 月为例进行详细介绍。

Frechet Distance 的时段运势轨迹相似性模型的初始数据是综合运势得分，可以通过 2.1.5 描述的方法获得，我在此不再赘述。之后是对获得的得分进行清洗。得分我们是通过 csv 文件读入的。首先对标题切片，分别一一对应储存好星座名称和日期，然后根据传入的星座名称选取对应的数据单独储存，原始数据类型为 str，运用 list 函数的内置函数 map 强制类型转换为 int，以字典对的形式存储到列表中，由于两种数据来源因反爬等各种因素，日期上并不是高度一致，因此我们需要通过循环进行检查，只选取相同日期的数据进行存储。

```
for i in range(0,n):
    x = temp[i]
    y = data_3[i]
    list9_Tarot.append((x,y))
list9_Tarot = list(reversed(list9_Tarot))
#print(list9_Tarot)
xingzuowu = []
k = 0
for j in range(0,nn):
    if list9_Tarot[j][0] == list9_xingzuowu[k][0]:
        xingzuowu.append(list9_xingzuowu[k])
    elif list9_Tarot[j][0] > list9_xingzuowu[k][0]:
        k = list9_Tarot[j][0]-1
        xingzuowu.append(list9_xingzuowu[k])
list9_xingzuowu = xingzuowu
```

图 14 时段运势轨迹相似性模型所需数据的清洗代码

2.2 各星座爱情运势分析与可视化

2.2.1 爱情运势文本词云展示

通过导入 cv2 中的 imread 模块，使词云展示呈现特定的图形。考虑到在这个模块我们研究的是爱情运势，我选定的图形为一颗大爱心。之后调用 wordcloud 库画出词云图。我在报告中展示的图片为天秤座 9 月的爱情运势。在词云中，出现次数越多的词句在图中显示的字体就越大，可以很好地看到，词云能在一定程度上反应星座的爱情运势。

```
pic = imread("E:/DaoLun/QiMo/img/1.png")#然后读出这个图片
w_2 = wordcloud.WordCloud(mask=pic,font_path = "E:/DaoLun/qimo/path/songti.ttf",width = 758,height = 659,background_color = "white")
w_2.generate(txt)
plt.imshow(w_2, interpolation="bilinear")
plt.title("天秤座爱情文本")
plt.rcParams['font.sans-serif']=['SimHei'] #显示中文
plt.show()
```

图 15 词云展示核心代码

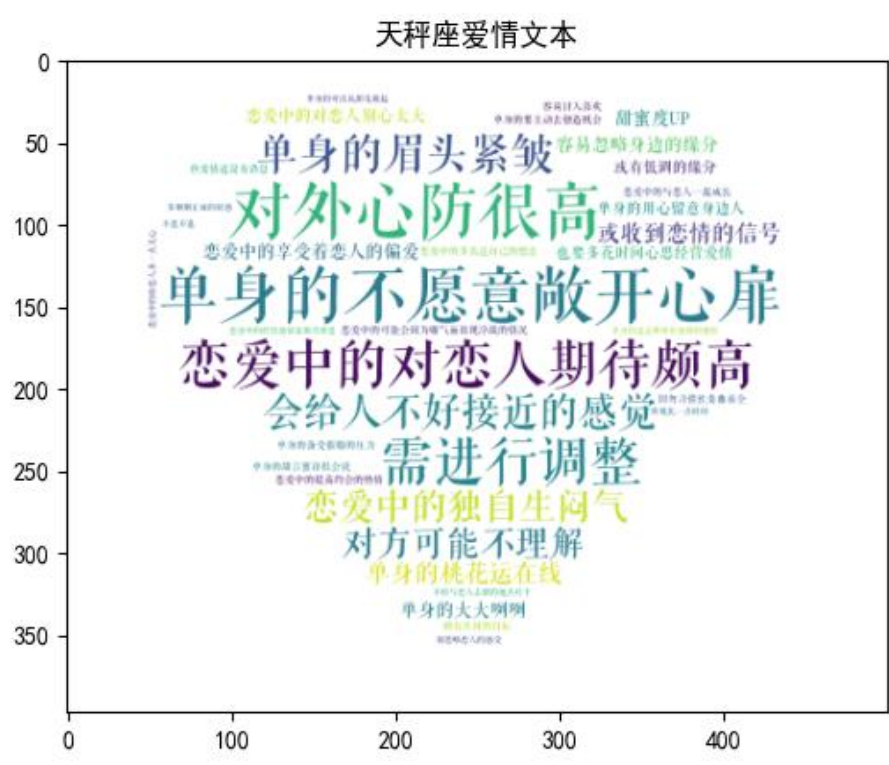


图 16 天秤座 9 月爱情运势文本词云展示

2.2.2 基于动态柱状图的契合星座判断

我的动态柱状图是用 html、css、js 三兄弟显示的。通过对 12 星座从 9 月 1

日到 12 月 20 日每日契合星座的出现次数进行动态展示，得出某个星座的最佳契合星座。关键代码是 js 文件的编写，我先在 echarts 上找到了一个实例，然后基于准备好的数据 dom 模块对这个实例进行初始化，读取第一个数据进行展示，然后编写了一个特定函数（图 17）更新后续数据，从而实现柱状图的动态变化。

```
// 更新数据
function updateYear(year)
{
    var cname = year.cname.split(',');
    var cut = year.cut.split(',');
    var n = cname.length;

    var index = [];
    for (let i = 0; i < n; ++i) {
        index.push(i);
    }
    index.sort((a, b) => {
        return Number(cut[a]) > Number(cut[b]) ? -1 : 1;
    });

    var yAxis = []
    var series = []
    for (let i = 0; i < n; ++i) {
        yAxis.push(cname[index[i]]);
        series.push(cut[index[i]]);
    }
    console.log({cname, cut, index})
    option.yAxis.data = yAxis;
    option.series[0].data = series;
    option.graphic.elements[0].style.text = year.cdate;
    // 使用刚指定的配置项和数据显示图表。
    console.log(option.yAxis.data);
    console.log(option.series[0].data);
    console.log(option.graphic.elements[0].style.text);
    myChart.setOption(option);
}
```

图 17 数据更新函数核心代码

由于文件格式限制，我无法在该报告中动态展示，特截取了金牛座的契合星座画面，具体的动态画面还请老师查看相应的网页。本次的动态柱状图我做了两版，一个的数据来源是星座屋网站，另一个的数据来源是闹闹女巫每日星运。纵坐标是 12 个星座，随意定格其中一个画面，展示的是某星座（画面右下角灰色字体）这一段时间的契合星座。比较两版柱状图的数据展示，我惊喜的发现，12 星座中有 8 个星座的契合星座是一致的。这一结果一定程度上符合我最开始提出的不同博主的星象分析应该总体趋于一致。

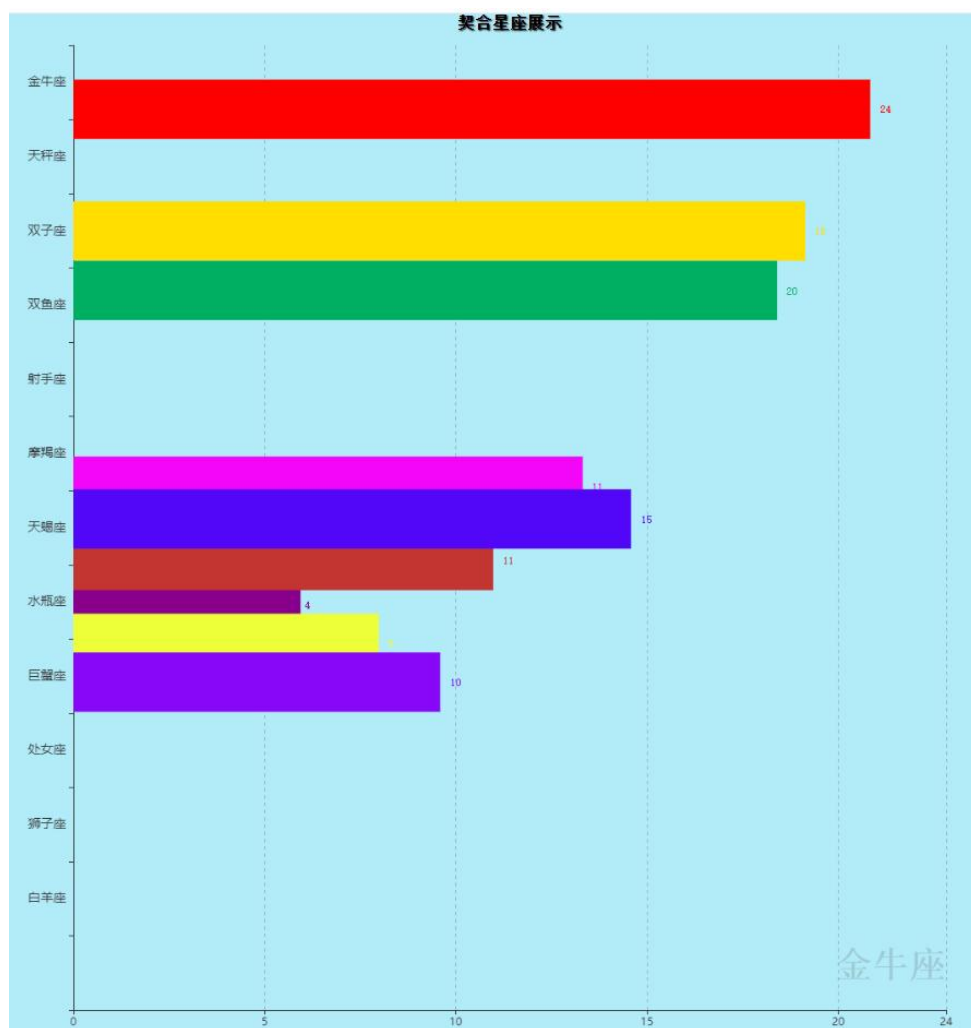


图 18 金牛座契合星座画面展示

2.2.3 基于每日爱情运势得分的图象模拟

在寻找曲线的波峰、波谷时，由于数据帧数多的原因，导致生成的曲线图噪声很大，不易寻找规律。为了降低噪声干扰，需要对曲线做平滑处理，让曲线过渡更平滑。常见的对曲线进行平滑处理的方法包括：Savitzky-Golay 滤波器、插值法等。插值，简单来说，就是通过构造多项式的最高次数，根据原有数据进行填充，最后生成的曲线一定过原有点。而拟合是通过原有数据，调整曲线系数，使得曲线与已知点集的差别（最小二乘）最小，最后生成的曲线不一定经过原有点。由于在数据绘制时，我并不是每天的数据都绘制，数据量有所降低，我采用的是 `make_interp_spline` 插值法。我在报告中展示了天秤座从 9 月 1 日到 12 月 20 日的爱情运势走向，可以看出整体是呈现下降趋势的。

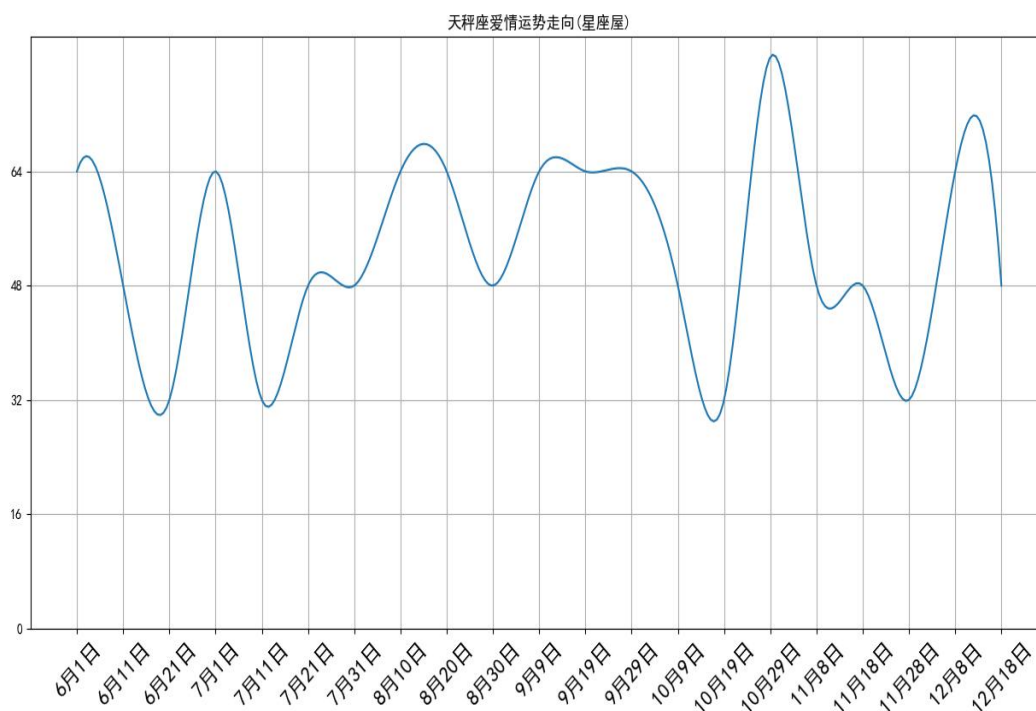


图 19 天秤座爱情运势走向展示

2.3 基于饼状图的幸运颜色拟合

通过 `matplotlib.pyplot` 库可以自由设置饼图的参数，描画出我们想要的模型。本次的数据来源是星座屋网站和闹闹女巫每日星运。内圈是星座屋网站，外圈是闹闹。原理与 2.2.2 契合星座的获得类似，在图中所占比例最大的即是该星座的契合星座。我下面展示的是白羊座的幸运颜色拟合。通过比对 12 个星座画出来的图片，我很遗憾的发现，没有任何一个星座两个博主分析的幸运颜色是拟合的。

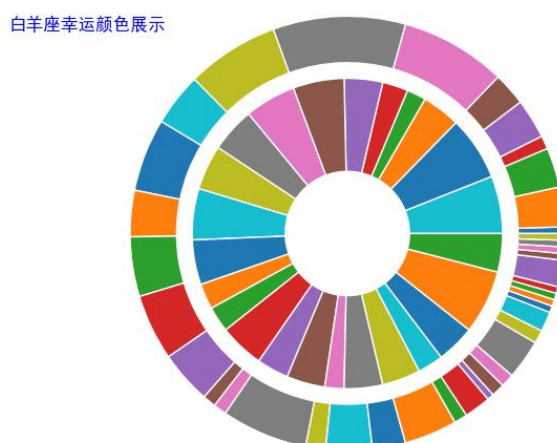


图 20 白羊座幸运颜色拟合展示

2.4 基于 Frechet Distance 的时段运势轨迹相似性模型

在 2.2.3 板块我做出了每日爱情运势得分的图像模拟，但是我只展示了一个数据源。其实，该图像的数据来源可以有两个：星座屋网站和星座女巫 Tarot；我们还可以做出事业学业的、健康运势的。那么对于同一星座同一类运势不同数据源所作出来的轨迹是否趋于一致呢？我们并不能通过肉眼主观判断，而应该建立科学的数学模型客观分析。

目前，判断两条轨迹的相似性方法有很多：基于点方法的有 EDR, LCSS, DTW 等；基于形状的方法有 Frechet, Hausdorff；基于分段的方法有 One Way Distance, LIP distance；基于特定任务的方法有 TRACCLUS, Road Network, grid 等。而我们本次模型构建过程中的数据源来自两个不同的博主，打分的偏好性不同，一个偏向于打高分，一个偏向于打低分。因而，我们并不能采用广泛使用的点方法简单计算距离，而是应采用基于形状的方法。更加通俗的讲，在图 21 中，我们可以看到两条轨迹的形状基本趋于一致，但是一个大，一个小，我们并不能因为大小问题断定两个轨迹不相似。我本次采用的是学术界更被普遍承认的 Frechet Distance。



图 21 轨迹举例

对于 Frechet Distance 定义，其中最为简单直观的一个理解，主人和狗在两条不同的轨迹上运动，主人和狗之间是由狗绳相连接的，Frechet Distance 即两者能各自走完整个轨迹的情况下满足条件的狗绳的最短长度。Frechet Distance 考虑到了沿曲线各点的位置和顺序，因而也比众所周知的 Hausdorff distance 更好。前面提到，目前对于 Frechet Distance 的论文研究并不是很多，基本停留在数学层面，与计算机的结合并不多。因而我选择了计算 Frechet Distance 的近似算法，该算法通过对曲线的无限近似逼近精确的 Frechet Distance。该算法的伪代码如图 22 所示。

此后我通过编写 python 代码复现了这个伪代码（算法代码复现真是我至今做过最痛苦的事情）。然后计算出了 12 星座在 9 月到 12 月这个时期内每个月不同运势的 Frechet Distance。从数据中我们可以看出，虽然是不同的星座，不同的月份，不同的运势，但是计算出来的 Frechet Distance 基本以 30 为中心，上下略微浮动，但总体是稳定趋于一致的，这无疑是令我开心和激动的。这验证了两个星座博主对于这三类运势的分析基本趋于一致！

```

Function dF( $P, Q$ ): real;
  input:    polygonal curves  $P = (u_1, \dots, u_p)$  and  $Q = (v_1, \dots, v_q)$ .
  return:   $\delta_{dF}(P, Q)$ 

   $ca$  : array [1.. $p$ , 1.. $q$ ] of real;
  function  $c(i, j)$ : real;
    begin
      if  $ca(i, j) > -1$  then return  $ca(i, j)$ 
      elseif  $i = 1$  and  $j = 1$  then  $ca(i, j) := d(u_1, v_1)$ 
      elseif  $i > 1$  and  $j = 1$  then  $ca(i, j) := \max\{c(i-1, 1), d(u_i, v_1)\}$ 
      elseif  $i = 1$  and  $j > 1$  then  $ca(i, j) := \max\{c(1, j-1), d(u_1, v_j)\}$ 
      elseif  $i > 1$  and  $j > 1$  then  $ca(i, j) :=$ 
         $\max\{\min(c(i-1, j), c(i-1, j-1), c(i, j-1)), d(u_i, v_j)\}$ 
      else  $ca(i, j) = \infty$ 
    return  $ca(i, j)$ ;
  end; /* function  $c$  */

  begin
    for  $i = 1$  to  $p$  do for  $j = 1$  to  $q$  do  $ca(i, j) := -1.0$ ;
  return  $c(p, q)$ ;
  end.

```

图 20 Frechet Distance 近似算法的伪代码

水瓶座.csv	
1	,月份,爱情运势,事业学业,财富运势
2	0,9,30.59411708155671,40.0,30.14962686336267
3	1,10,31.622776601683796,30.265491900843116,30.066592756745816
4	2,11,30.59411708155671,34.9857113690718,30.59411708155671
5	3,12,45.0,50.0,40.0
6	4,9_12,45.0,50.0,40.0
7	
白羊座.csv	
1	,月份,爱情运势,事业学业,财富运势
2	0,9,30.265491900843116,45.0,50.0
3	1,10,30.0,45.0,30.066592756745816
4	2,11,45.0,30.01666203960727,45.0
5	3,12,30.14962686336267,30.59411708155671,31.622776601683796
6	4,9_12,30.413812651491103,45.0,50.0

图 21 水瓶座和白羊座的 Frechet Distance 数值

为了更加直观的分析得出的 Frechet Distance 数值，我选择从三个不同的维度：不同运势、不同月份、不同星座，描画出与之相关的气泡图。

在图 22（不同星座对比）中，我们发现整体气泡颜色以青色为主，说明两个博主在财富运势的分析趋向性较低，相反紫色和橙色的占比则较少，说明两个博主在爱情运势和事业学业运势上的分析基本趋于一致。

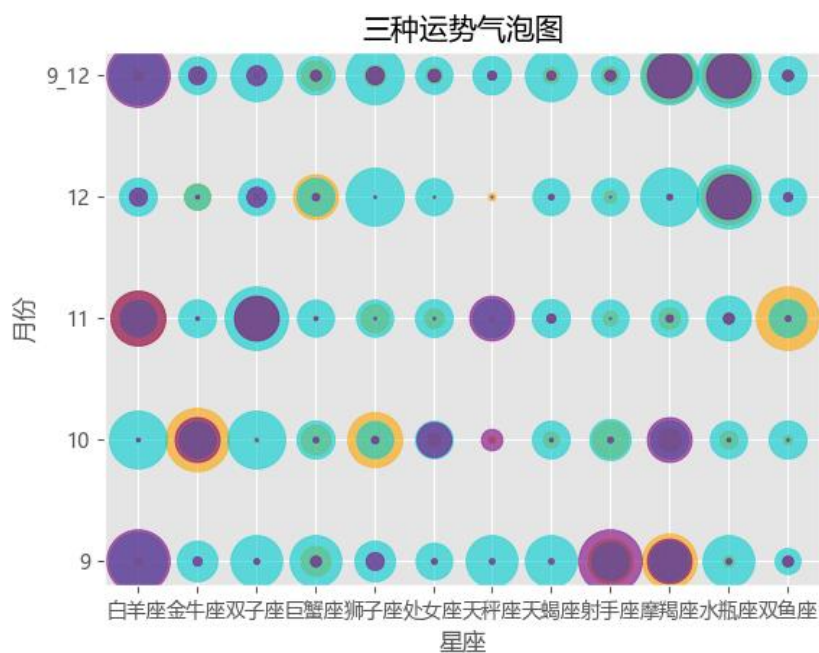


图 22 不同运势维度看 Frechet Distance

在图 23（不同月份对比）中，我们可以看到总体气泡大小相差不大，几乎看不到同一个气泡不同颜色的交替，这说明两个博主在不同月份上的分析是高度一致的。在该图中还有一个意料之外的有趣发现。我们发现，两个临近星座的颜色是极其相似甚至完全一样的。比如说金牛座和白羊座都是红色，这并不是我在代码编写过程中刻意规定的，而是根据数据特点自然呈现的。这一现象正好印证了星座圈对于“临近星座”的分析，比如说，金牛座和白羊座就是临近星座。这无疑是一个可喜可贺的意外收获。

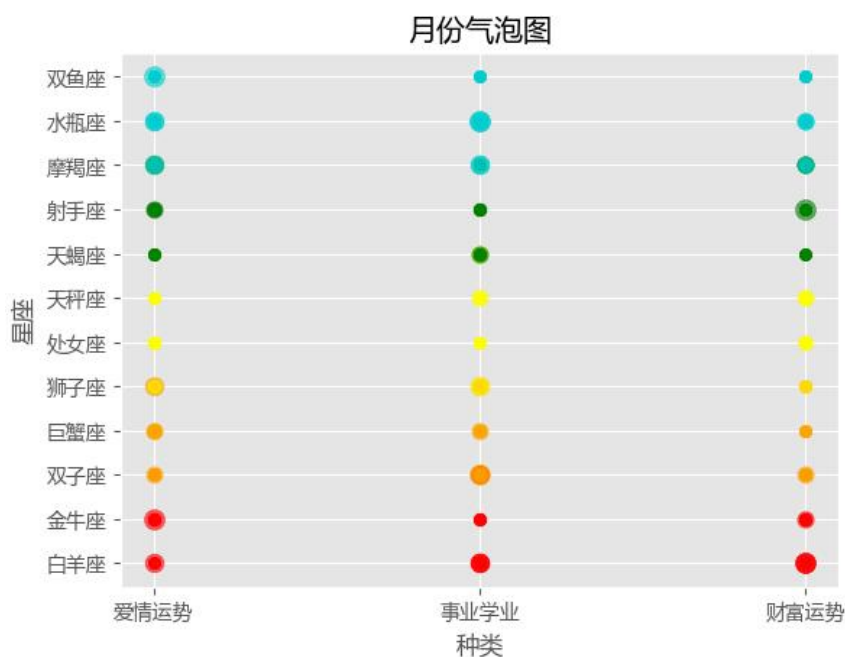


图 23 不同月份维度看 Frechet Distance

在图 24（不同星座对比）中，我们可以看到气泡交叠数量以两个为主，说明两位博主对 12 星座运势分析差不多有 10 个都是趋于一致的！这种高度一致性令深夜做出图的我倍感精神！具体到坐标分析，在 11 月份的爱情运势位置，气泡呈现了不同颜色的多种交叠，说明两位博主对这一运势的 12 星座分析是不同的，但是这种不同出现的频率并不高！

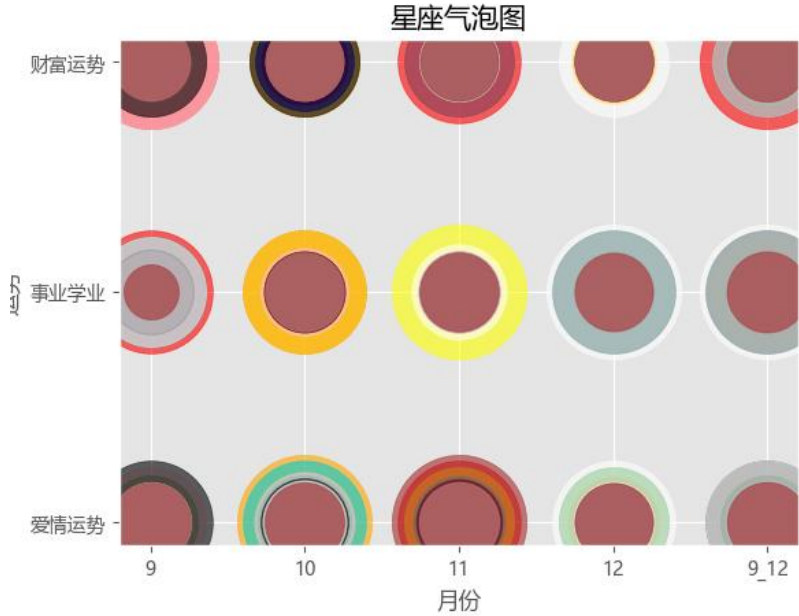


图 24 不同星座维度看 Frechet Distance

2.5 基于幸运数字的拟合优度模型

每日运势分析中最常见的内容就是该日幸运数字，在数学上，1 和 2 可能是完全不同的，但是在星座圈中，数字的更迭是有规律的，1 和 2 是相似的概念，数字的相似性随两个数字差的增大而增大。那么我们可以通过分析不同数据来源对每日幸运数字的分析的拟合来判断不同博主的分析一致性。但是该种数据类型的拟合和 2.3 是完全不同的，因而我们需要重新选择一种数学模型：拟合优度模型。此外，我选择的数据来源是星座屋网站和闹闹每日星运公众号。

设 y 为待拟合数据， y 的均值为 \bar{y} ，拟合数据为 \hat{y} ，则：

1. 总平方和 SST(total sum of squares) : $\sum_{i=1}^n (y_i - \bar{y})^2$
2. 回归平方和 SSR(regression sum of squares) : $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
3. 残差平方和 SSE(error sum of squares) : $\sum_{i=1}^n (y_i - \hat{y}_i)^2$

确定系数:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

图 25 拟合优度计算公式

拟合优度（Goodness of Fit）是指回归曲线对观测值的拟合程度。度量拟合优度的统计量是可决系数（亦称确定系数） R^2 。 R^2 最大值为 1。 R^2 的值越接近 1，说明回归曲线对观测值的拟合程度越好；反之， R^2 的值越小，说明回归曲线对观测值的拟合程度越差。总而言之，拟合优度是用于度量拟合曲线对于原始数据拟合效果的好坏，拟合优度 R^2 越接近 1 说明拟合优度越好，一般来说，拟合优度到达 0.7 以上就可以说拟合效果不错了。计算方法如图 25 所示，复现代码如图 26 所示。

```
def __ssr(y_fitting, y_no_fitting):
    """
    计算SSR(regression sum of squares) 回归平方和
    :param y_fitting: List[int] or array[int] 拟合好的y值
    :param y_no_fitting: List[int] or array[int] 待拟合y值
    :return: 回归平方和SSR
    """
    y_mean = sum(y_no_fitting) / len(y_no_fitting)
    s_list = [(y - y_mean)**2 for y in y_fitting]
    ssr = sum(s_list)
    return ssr

def __sse(y_fitting, y_no_fitting):
    """
    计算SSE(error sum of squares) 残差平方和
    :param y_fitting: List[int] or array[int] 拟合好的y值
    :param y_no_fitting: List[int] or array[int] 待拟合y值
    :return: 残差平方和SSE
    """
    s_list = [(y_fitting[i] - y_no_fitting[i])**2 for i in range(len(y_fitting))]
    sse = sum(s_list)
    return sse

def goodness_of_fit(y_fitting, y_no_fitting):
    """
    计算拟合优度R^2
    :param y_fitting: List[int] or array[int] 拟合好的y值
    :param y_no_fitting: List[int] or array[int] 待拟合y值
    :return: 拟合优度R^2
    """
    SSR = __ssr(y_fitting, y_no_fitting)
    SST = __sst(y_no_fitting)
    rr = SSR / SST
    return rr
```

图 26 拟合优度代码复现

我先用这些折线图画出了 12 星座 4 个不同月份两个博主分析的每日幸运数字，然后针对每一张图计算出了它的拟合优度。图 27 显示了天秤座 9 月份的幸运数字分析。总览计算出的所有拟合优度（图 28），我们可以看到，大部分数据大于 0.7 小于 1，间或有低于 0.5 的欠拟合数据和高于 1 的过拟合数据，可以说，两个博主的分析的拟合效果是非常好的。之后我又通过可视化手段综合分析了所有计算出的拟合优度。在图 29 中，整体来看，气泡大小趋于一致，偶尔出

现过大过小的数据，符合前面对数字的直接分析。从横坐标的角度来看，每个星座在这 4 个月份中因拟合优度太低而导致的气泡过小甚至不显示只出现一次，说明大约只有 1 个月两个博主的分析是完全不同的；从纵坐标的角度来看，12 月份的气泡大小是变化不定的，说明两个博主在该月的幸运数字分析相似度不高，但是其他月份的相似度还是比较合理的。

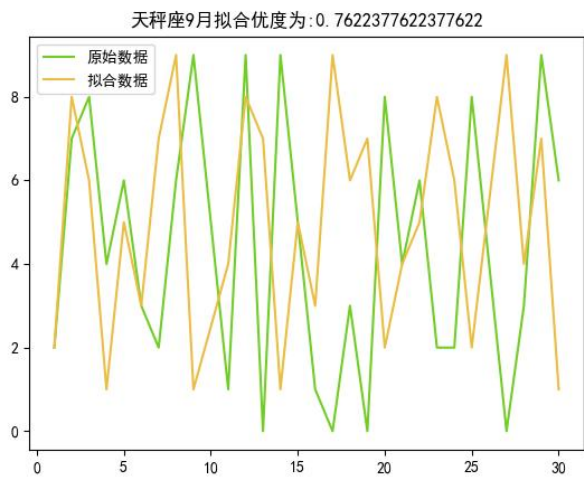


图 27 天秤座 9 月份幸运数字

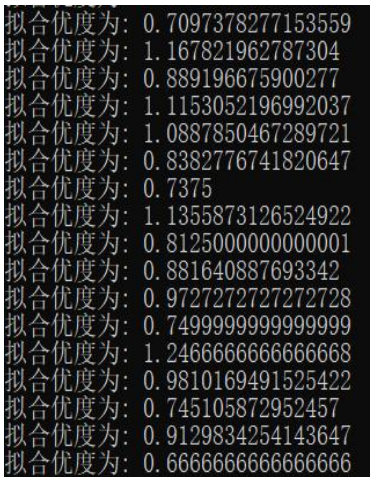


图 28 拟合优度部分数据

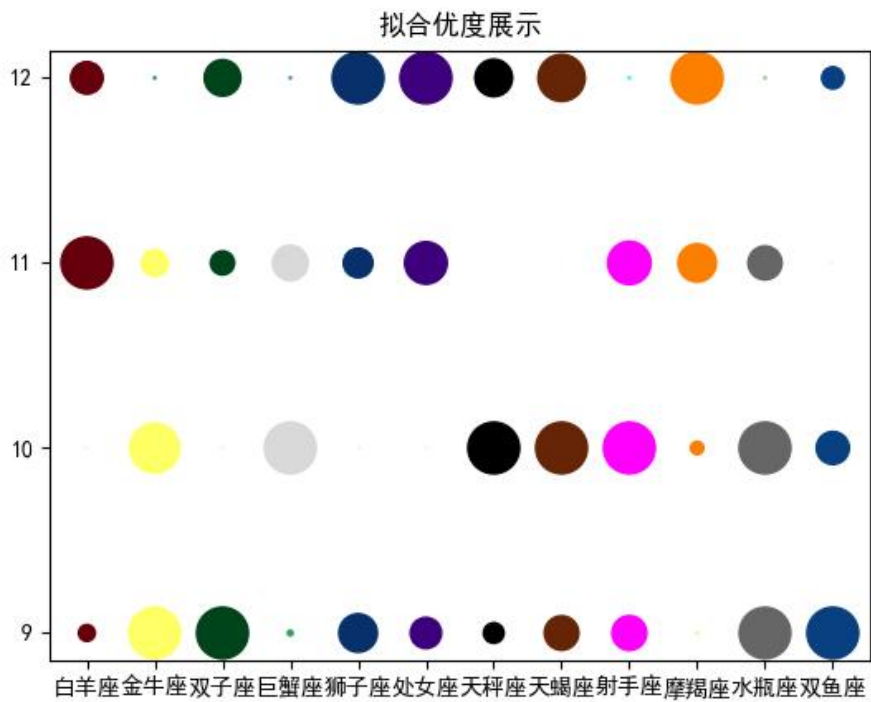


图 29 拟合优度综合展示

3. 结论与展望

本次实验大致一共分为两个部分。各星座爱情运势分析与可视化是第一部分，我们从中可以了解到从 2022 年 9 月 1 日到 2022 年 12 月 20 日不同星座的爱情运势主旋律，最佳匹配星座和爱情运势变化走向。拟合是本次实验的第二部分，也是做重要的部分。第一个幸运颜色的拟合宣告失败，不同博主分析的幸运颜色完全不同。接着，基于 Frechet Distance 的时段运势轨迹相似性模型分析了 12 星座不同博主分析的各种运势的变化轨迹的相似性。通过 2.4 的分析，我认为在该领域，选取的两个博主的分析基本趋于一致。最后，基于幸运数字的拟合优度模型分析了 12 星座不同博主分析的每日幸运数字的相似性，通过 2.5 的分析，我认为在该领域，选取的两个博主的分析也是趋于一致的。综合本次实验的所有内容，我们可以得出不同博主对于同一星相的分析总体上看是趋于一致的，从这个角度来看，当下的星座分析并不是博主凭空捏造的，星座分析具有一定的可信度。

但是星座分析可信并不代表我们要盲从。今日某个博主分析到水瓶座的学业运势节节高升，那么该星座的学生今天就可以停笔弃本，等待幸运的降临吗？那必然是不行的。首先，星座预测是基于大数定律的综合预判，对于大多数人来说有一定的准确性，但不是人人都应验。另外，通过前面的分析，我们可以知道，不同博主的分析并不是完完全全一模一样的，在某个月份某类运势上可能就是相互背离的，在这种情况下，我们又怎么能断定，某个博主的分析一定是对的呢。我认为，星座分析更多的是一种提醒，今日预测到我的星座容易出现交通事故，那么我在出行的时候便会多加小心，从而免除灾祸。星座的提醒作用可能就是俗语“小心驶得万年船”的最佳诠释了。

在本此实验中我只分析了三个博主的数据，若有感兴趣的读者可以选取其他博主的分析，更加全面的比对。我爬取到的数据还有一大块值得发掘的宝藏：每日各类运势的文本情感倾向性分析。如果能够攻下这个难题，相信本次实验所做的论证会更加具有说服力。对星座感兴趣的读者还可以继续分析不同博主帖子下的评论分析，通过受众的普遍反映，既能看出某个星座的当日真实运势，也能判断某个博主的分析是否具有可信度。

4. 前期未成形模型

助教老师在课上讲解的时候，提到注重工作过程，即使最后结果不尽如人意也没关系。所以我在此讲解一下我前期未成形的工作。如果我的工时用“1”表示，那么前期的时长达到甚至高于了“1/2”，我最开始想做的是爬取抖音上各大星座博主的视频和评论，通过对博主视频内容的分析研究推算出各大星座的性格模型，通过网友的评论反馈推算出判断博主分析准确程度的模型。众所周知，抖音是目前各大平台里反爬机制做的最好的平台。我花费了大量的时间去研究抖音的反爬机制，去学习相关的爬虫代码。比较成功的是我爬取各大博主视频的代码很成功，百发百中，但是爬取评论的代码不尽如人意，每次的爬取量不能确定，多的时候有 4000 多条，少的时候只有 10 多条，大多数情况下一条也爬取不下来。之后，我调用北大的一个语音转文本的开源软件包 Buzz，将视频内容解析出来。但是无奈在抖音上的数据量还是太少，无法进行更加深入的分析，迫不得已在后期紧急换方向，这也是我本次实验过程的一大遗憾吧。


```


douyin.py > ...
24
25 def http_get(url):
26     resp = requests.get(url, headers=headers)
27     return resp
28
29
30 def get_item_id(short_url):
31     res = http_get(short_url)
32     item_id = re.findall(r"(?<=video/)\d+", res)
33     print(item_id)
34     return item_id
35
36
37 def get_play_url(item_id):
38     api_url = "https://www.iesdouyin.com/webapi/v1/feed/item/"
39     api = http_get(api_url).text
40     api = json.loads(api)
41     playwm = api['item_list'][0]['video']['playwm']
42     play = playwm.replace('/playwm/', '/playwm/')
43     return play
44
45
46 def fetch(short_url):
47     item_id = get_item_id(short_url)
48     return get_play_url(item_id)
49
50
51 url = fetch("https://v.douyin.com/hDd3u84/")
52 print(url)

```

图 30 抖音视频爬取核心代码

G	H	I	J	K	L	M	N
评论人	评论时间	评论内容	点赞数	二级评论人	二级评论时间	二级评论内容	二级评论点赞数
0***	2022.03.18	我是水瓶座，不喜欢忽冷忽热，就喜欢直接明显的爱意，还有偏爱，					
0***	2022.03.18	我是水瓶座	2038	再***	2022.03.18	对对对就是这样没有愿	
0***	2022.03.18	我是水瓶座	2038	闭***	2022.03.18	芽曜掉没啦	
0***	2022.03.18	我是水瓶座	2038	h***	2022.03.18	懂	
0***	2022.03.18	我是水瓶座	2038	0***	2022.03.18	我也是	
0***	2022.03.18	我是水瓶座	2038	0***	2022.03.18	感觉不到明显的爱意直	
0***	2022.03.18	我是水瓶座	2038	再***	2022.03.18	我把他逼走了	
0***	2022.03.18	我是水瓶座	2038	0***	2022.03.18	那是他的损失，他没有	
0***	2022.03.18	我是水瓶座	2038	再***	2022.03.18	是我的过错	
0***	2022.03.18	我是水瓶座	2038	0***	2022.03.18	爱你的人会回来找你的	
0***	2022.03.18	我是水瓶座	2038	再***	2022.03.18	那可能就是不爱了吧	
0***	2022.03.18	我是水瓶座	2038	0***	2022.03.18	再等等，给他点时间，	
0***	2022.03.18	我是水瓶座	2038	再***	2022.03.18	都放个假吧，后面的事	
0***	2022.03.18	我是水瓶座	2038	[***	2022.03.24	@Gemini	
0***	2022.03.18	我是水瓶座	2038	e***	2022.03.24		
0***	2022.03.18	我是水瓶座	2038	天***	2022.04.07	是的	
0***	2022.03.18	我是水瓶座	2038	言***	2022.04.14	昊昊的崽崽 你看 水	
0***	2022.03.18	我是水瓶座	2038	昊***	2022.04.14	对方知道水瓶的心意但	
0***	2022.03.18	我是水瓶座	2038	七***	2022.05.31	@刺猬不刺	

图 31 抖音视频下爬取到的评论

 tiancheng.txt - 记事本

文件(E) 编辑(E) 格式(O) 查看(V) 帮助(H)

大家好,我是老白呀,今天我们来飞机一下 天称的理想型是什么样子的 天
球 给天称做很 一种气洞感 他就能去了 打了很感应 然后打游戏 打

图 32 视频语音转文本成果展示

5.致谢

长大以后，一直在感慨，时间过的好快，一眨眼的时间，本学期的学习就结束了。我在未来的某个时刻一定会无比怀念周一王老师在课堂上讲解相关知识时那特别磁性悦耳的声音，周二助教老师在教室里给我讲代码时那认真可爱的面庞！非常感谢王伟老师和两位助教老师对我的谆谆教诲与鼎力相助！在本次期末实验过程中，我受新冠病毒的影响较深，15天才转阴，三位老师非常体谅我的难处，同意我延期答辩，我不胜感激。

在我们步入新冠第四年的时候，生活开始逐步回归常态。大面积的“暂停时光”已是过去式，“一刀切”的防疫封控终于结束了！相信下学期我们都会度过一个美丽的春天，欣赏丽娃河畔的杨柳依依！感谢与我一路同行的老师和同学，敬祝安好！

6.引用

[1] Eiter T, Mannila H. Computing discrete Fréchet distance[J]. 1994.

[2] 赵松山. 对拟合优度 R^2 的影响因素分析与评价[J]. 东北财经大学学报, 2003 (3): 56-58.