微博个性化推荐系统

成员以及分工

成员	任务分配
彭高鹏（1120202579）	完成系统设计，编写文档
孙瑞祥（1120201870）	收集数据，协助算法实现，编写文档
张雲赫（1120200951）	协同过滤算法实现，编写文档
马梓奥（1120202685）	基于内容推荐算法实现，编写文档
李牧宇（1120200800）	编写文档

一、问题提出

在当今时代，社交媒体平台已经成为人们获取信息、交流观点、分享生活的重要途径之一。微博作为其中一种重要的社交媒体平台，每天都会涌现出海量的信息，随着用户数量的增加和信息量的爆炸式增长，如何为每位用户提供个性化的微博推荐成为了一项重要的挑战。本项目旨在利用数据挖掘技术，基于用户历史行为、兴趣等信息，构建一个个性化微博推荐系统，以提升用户体验，增加用户粘性，这能有效提高微博平台的活跃度和用户满意度。

二、数据获取和分析

1. 数据获取及预处理

本项目通过微博API获取了两类关键数据：用户的推文数据和关注数据。这些数据被用来分析用户互动和关注特征，以便理解用户行为和优化社交媒体策略。

1.1 数据来源

- **推文数据**：来自微博用户的推文记录，包含了互动数（点赞、评论、转发）和推文内容。
- **关注列表数据**：包含了用户的关注者信息，例如性别和是否互相关注。

1.2 数据说明

- **推文数据**：包括了字段如`created_at`（创建时间）、`text`（文本内容）、`reposts_count`（转发数）、`comments_count`（评论数）和`attitudes_count`（点赞数）。
- **关注者数据**：字段包括`id`、`screen_name`（昵称）、`gender`（性别）、`followers_count`（关注者数量）、以及`follow_me`（是否关注我）。

1.3 数据预处理

1 时间数据处理

- **日期时间转换**：推文的`created_at`字段原始格式为字符串。我们使用Python的`pandas`库将这些字符串转换成`datetime`对象，使其可以进行时间序列分析。

2 文本数据清洗

- **文本清洗**：对推文内容`text`进行了基本的清洗，包括去除多余的空格、特殊字符和网页链接。这一步骤有助于后续的文本分析和特征提取。

3 缺失值处理

- **缺失值检查与处理**：我们检查了数据集中的缺失值，特别是关键的互动指标（点赞、评论、转发数）。对于少数缺失的数据，我们选择用平均值填充，以保持数据完整性。

2. 关注者数据预处理

1 数据类型转换

- **关注者数量转换**：关注者数据中的`followers_count`字段包含有"万"字的字符串表示（例如"46万"）。我们编写了函数将这些字符串转换为实际的数值类型，例如将"46万"转换为46000，以便进行数值运算和统计分析。

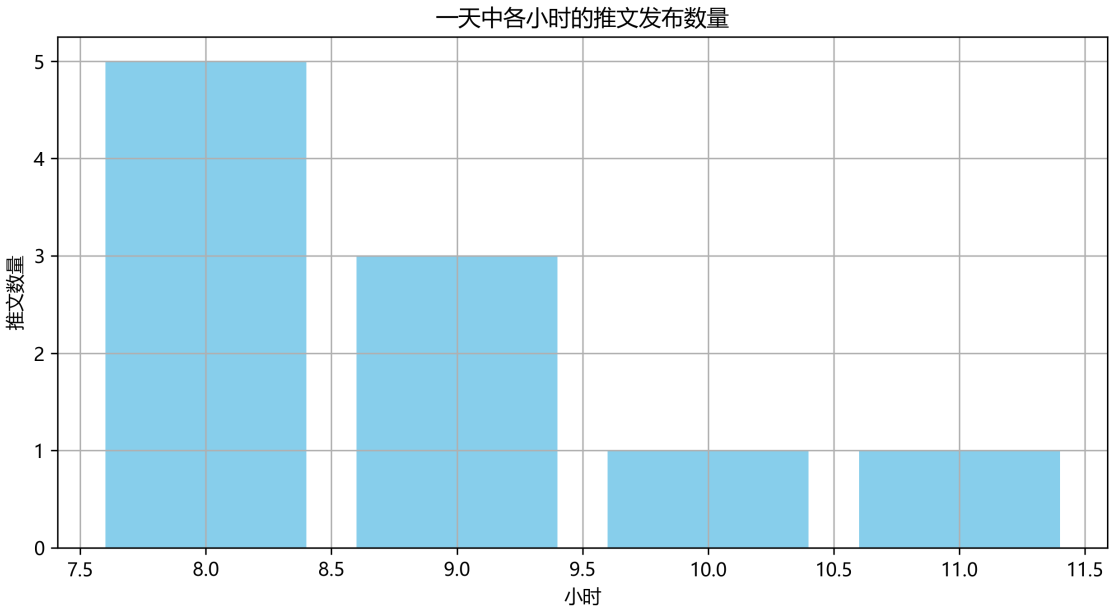
2 缺失值和一致性检查

- **性别字段处理**：在关注者数据中，性别字段`gender`可能存在缺失或不一致的情况。我们检查了该字段，对缺失的性别数据标记为"未知"，确保所有记录都有一致的性别标识。
- **互动状态检查**：检查了`follow_me`字段，确认是否有逻辑错误或缺失，并保证了其逻辑一致性。

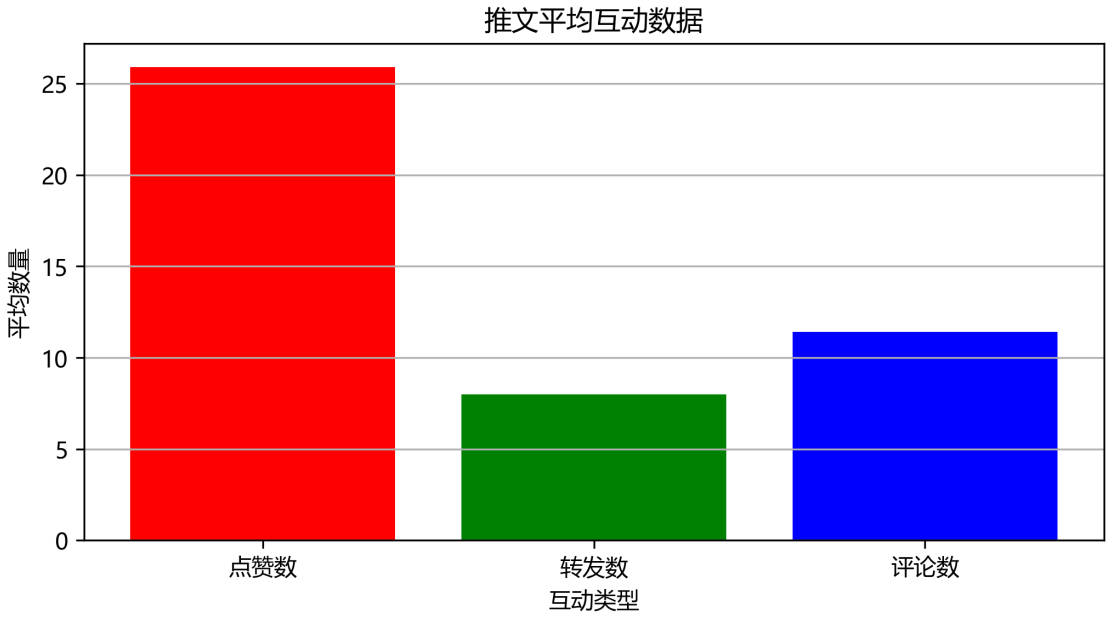
2. 数据分析与可视化

2.1 推文数据分析

- **活动时间分析**：分析了用户推文的时间分布，发现用户在某些特定时间段内更活跃。这有助于了解用户的日常使用习惯和最佳发布时间。

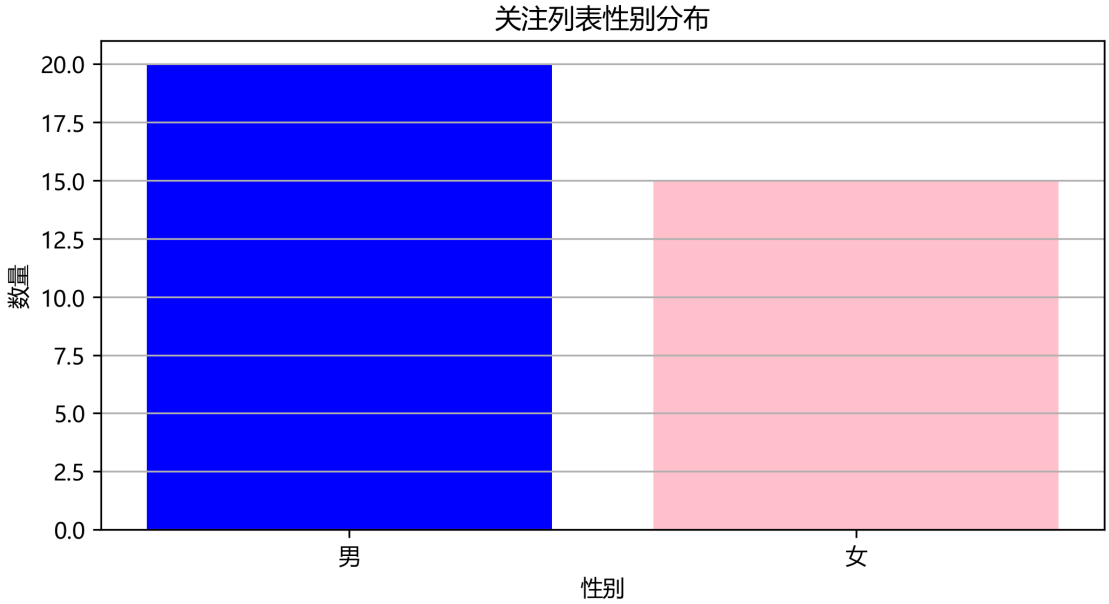


- **互动数据分析**：计算了推文的平均点赞数、评论数和转发数。这些数据提供了对哪些类型的内容更能激起用户互动的洞察。



2.2 关注数据分析

- **性别分布**：显示了关注列表中的性别比例，女性用户在某些样本中占多数。这种视觉化有助于快速识别用户关注列表的性别构成。



- **互动情况**：分析了关注者是否实际关注了该用户，发现大多数关注者并未互相关注，这可能影响用户的互动率和内容策略。

三、模型探索

3.1协同过滤算法

3.1.1模型介绍

基于用户的协同过滤模型（User-based Collaborative Filtering）是一种传统的推荐系统方法

它基于一个假设：

> 如果两个用户在过去对某些项的评价相似，那么他们在未来对其他项的评价也会相似。

该模型通常包括以下步骤：

- **相似度计算**：首先计算用户之间的相似度，常用的相似度度量包括皮尔逊相关系数、余弦相似度和杰卡德相似度等。

- **邻居选择**：根据相似度，为每个用户选择一组“邻居”（即相似的用户）。
- **生成推荐**：利用这些邻居的评分来预测目标用户对未知项的评分，推荐预测评分最高的项。

用户协同过滤的核心是计算用户之间的相似度和预测评分。下面是一些基本的公式：

- **用户相似度计算**（使用皮尔逊相关系数）：

  $$\text{sim}(u,v) = \frac{\sum_{i \in I}(R_{u,i} - \overline{R}_u)(R_{v,i} - \overline{R}_v)}{\sqrt{\sum_{i \in I}(R_{u,i} - \overline{R}_u)^2}\sqrt{\sum_{i \in I}(R_{v,i} - \overline{R}_v)^2}}$$

  其中，$R_{u,i}$ 是用户 u 对物品 i 的评分，$\overline{R}_u$是用户 u 的平均评分，I 是用户 u 和 v 都评分的物品集合。

- **预测评分**：

  $$\hat{R}_{u,i} = \overline{R}_u + \frac{\sum_{v \in U}(R_{v,i} - \overline{R}_v) \cdot \text{sim}(u,v)}{\sum_{v \in U} |\text{sim}(u,v)|}$$

  其中，U 是与用户 u 相似的用户集合。

3.1.2选择依据

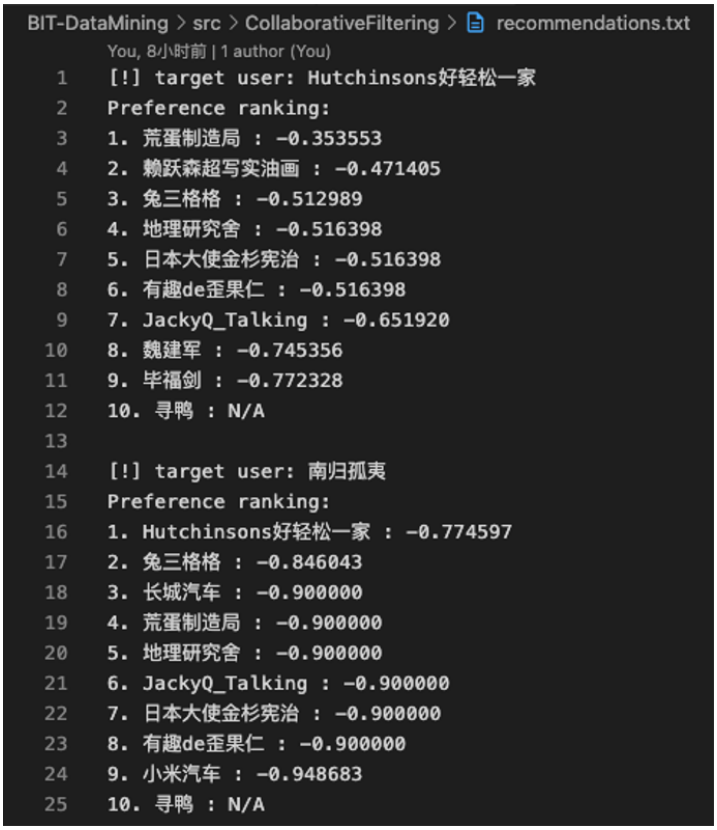
- 对于微博内容推荐系统的构建，采用基于用户的协同过滤模型直观且易于实现
- 微博用户的互动数据丰富，用户协同过滤能够有效利用这些数据来识别并推荐用户可能感兴趣的内容。

3.1.3实现过程

- 利用用户间的关注列表来发现潜在的兴趣相似性。
- 假设每个用户对他们关注的用户的“评分”为1，对不关注的为0。
- 计算用户间皮尔逊相关系数并据此排序，为用户推荐潜在兴趣的微博内容。

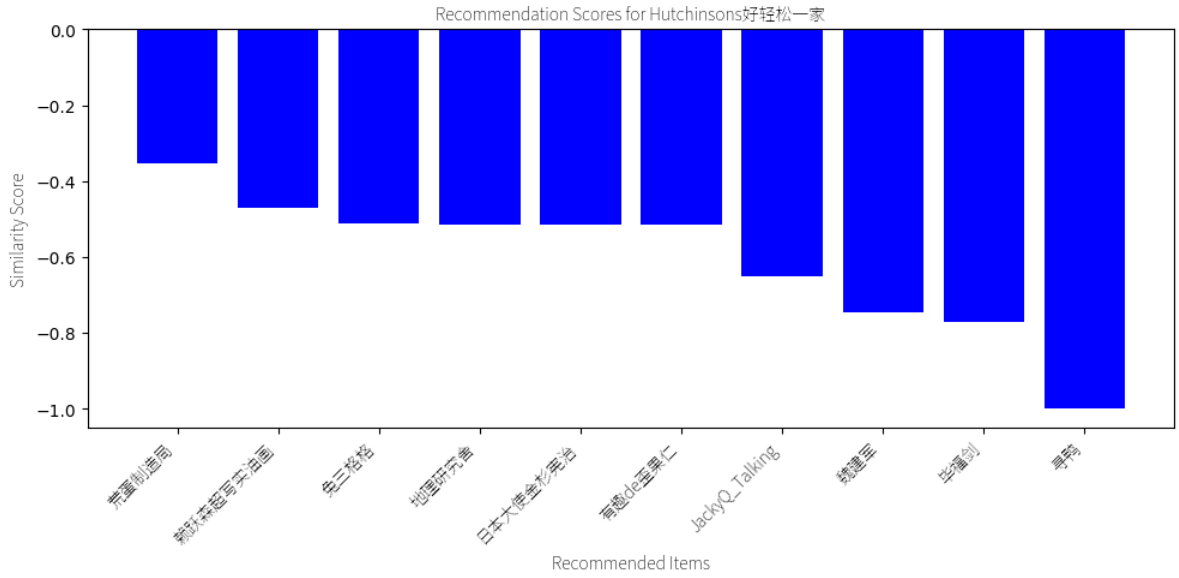
3.1.4结果展示

对于每个用户计算和其他用户对相似度分数，按照分数从高到低进行排序，将结果写入txt文件中。



结果中相似度分数越高代表用户越有可能对其所发表的微博内容感兴趣。

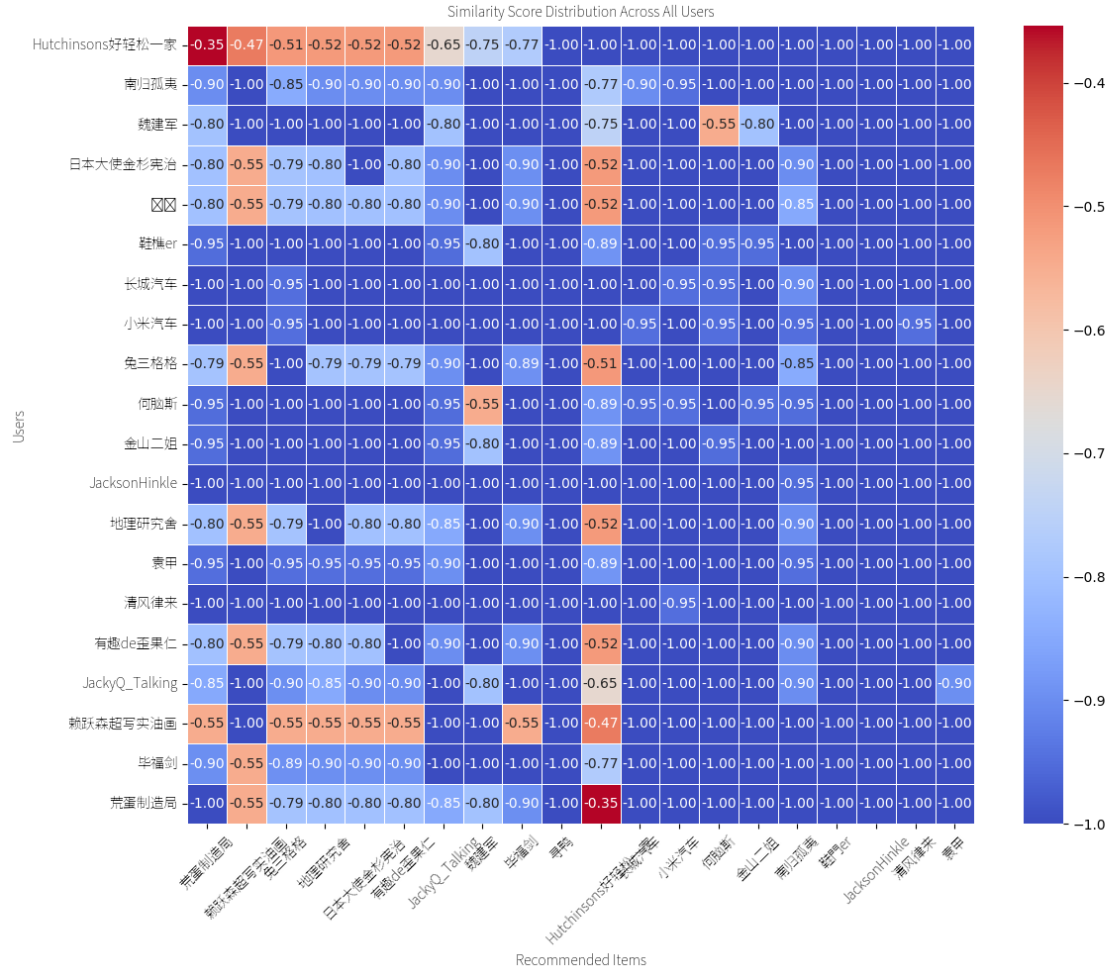
以微博用户“Hutchinsons好轻松一家”为例，分析其潜在兴趣：



用户“Hutchinsons好轻松一家”与微博用户“荒蛋制造局”相似度分数最高，因此最有可能对该用户的微博感兴趣。

用户“Hutchinsons好轻松一家”与微博用户“寻鸭”相似度分数最低，这是由于两个用户之间关注列表重合数量低。

进一步绘制热度图查看用户相似度分数分布情况。由于用户共同关注的账号数量不多，所以相似度分数分布较为分散。



3.2基于内容的推荐算法

3.2.1 算法介绍

基于内容推荐的原理侧重于分析和理解项目、内容本身的特征，以及用户历史行为对这些特征的偏好。推荐系统通过分析内容属性（如文字、图像、视频等）和用户的互动行为（如浏览、点赞、收藏等），学习用户的兴趣模型，并在此基础上向用户推荐内容。基于内容推荐系统使用各种机器学习算法来分析、计算内容之间的相似度和用户的预期兴趣，以提供个性化的内容推荐。

基于内容的推荐算法主要分为以下过程：

1. **内容分析（Content Analysis）：** 这是内容推荐系统的第一步，旨在通过算法（如TF-IDF，LDA，doc2vec等）处理和理解内容的特征。例如，TF-IDF 用于评估一字词对文档集或一个语料库中其所在文件的重要性，LDA 用于主题识别，doc2vec 用以生成文本内容的向量表示。
2. **用户画像构建（User Profile Creation）：** 系统通过分析用户的浏览历史、点赞、收藏等操作行为，结合其个人信息（如年龄、性别、地理位置等），创建出反映用户偏好的模型或画像。
3. **内容相似度计算：** 根据内容分析得到的特征，利用相似度计算方法（例如余弦相似度）来评估内容之间的相似程度。此步骤确保推荐的内容与用户已表现出兴趣的内容相似。
4. **个性化推荐：** 通过上述步骤，生成用户偏好与内容特征之间的映射，并利用这一映射关系进行个性化内容推荐。系统会根据用户画像和内容的匹配度来进行排序和推荐，从而提供个性化服务。

3.2.2 选择依据

本项目目标是构建一个个性化微博推荐系统，考虑到微博内容的丰富性（如文本、图片等）和用户行为数据的可用性，基于内容推荐在这个应用案例中有几点显著优势：

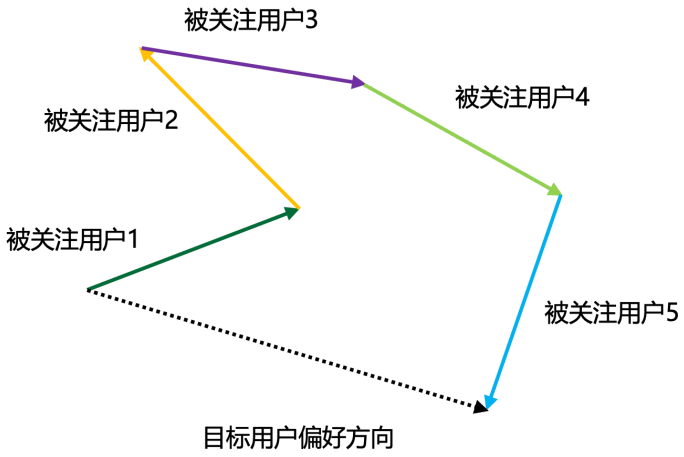
- **准确性：** 因为基于用户的历史行为和具体的内容特征，推荐通常较为精准。
- **透明度：** 推荐的原因（哪些内容特征）对用户来说较为明显，容易理解。
- **无需用户交互数据：** 初期推荐不依赖于用户之间的交互数据，容易实现冷启动。
- **内容理解：** 微博内容种类繁多，基于内容的推荐方法如TF-IDF、LDA、doc2vec可以有效地提取文本特征，理解内容的本质，为精准推荐提供强有力的支撑。
- **个性化体验：** 利用用户信息和操作行为构建用户画像，能够更好地捕捉到用户的个性化需求，从而优化用户在推荐系统中的个性化体验。

3.2.3 实现过程

- 内容分析：
    - 特征数据采集：从由微博爬取的结构化信息中提取有意义的自然语言文本，作为该内容的特征代理数据；
    - 特征提取算法：选用在中文语料上预训练的GTE模型（通用文本嵌入模型）对特征数据进行编码，得到内容对应的特征向量（经标准化得到的单位向量）；

Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (35 datasets)	Classification (9 datasets)	Clustering (4 datasets)	Pair Classification (2 datasets)
gte-large-zh	0.65	1024	512	**66.72**	71.34	53.07	81.14
gte-base-zh	0.20	768	512	65.92	71.26	53.86	80.44
stella-large-zh-v2	0.65	1024	1024	65.13	69.05	49.16	82.68
stella-large-zh	0.65	1024	1024	64.54	67.62	48.65	78.72

- 用户画像构建：
  - 用户偏好数据选取：以用户关注列表表征用户偏好；
  - 画像构建方式：以目标用户关注列表中所有被关注用户的特征向量（在内容分析一步中获得并保存）之和经标准化得到的向量作为用户画像；



- 相似度计算 & 个性化推荐：
  - 对于每个目标用户，计算其用户偏好向量与所有用户的内容特征向量的余弦相似度，按照分数从高到低进行排序；
  - 相似度分数越高代表目标用户越有可能对该被推荐用户所发表的微博内容感兴趣。

3.2.4 结果展示

对于每个用户，计算其画像和其他用户特征向量的余弦相似度分数，按照分数从高到低进行排序，排除用户已关注的用户，得到为该用户推荐的感兴趣用户。

```
[!] target user: JacksonHinkle
Preference ranking:
1. 专业戳轮胎熊律师: 0.9536
2. 番茄殿下: 0.9522
3. 伊能静: 0.9503
4. 一个专员: 0.9502
5. 并不软的软喵子: 0.9476
6. 咕咕咕_乐: 0.9465
7. CyberZhiqi: 0.9449
8. 铁手叫兽: 0.9433
9. 商建刚: 0.9414
10. 三联生活周刊: 0.9409
```

四、系统设计和展示

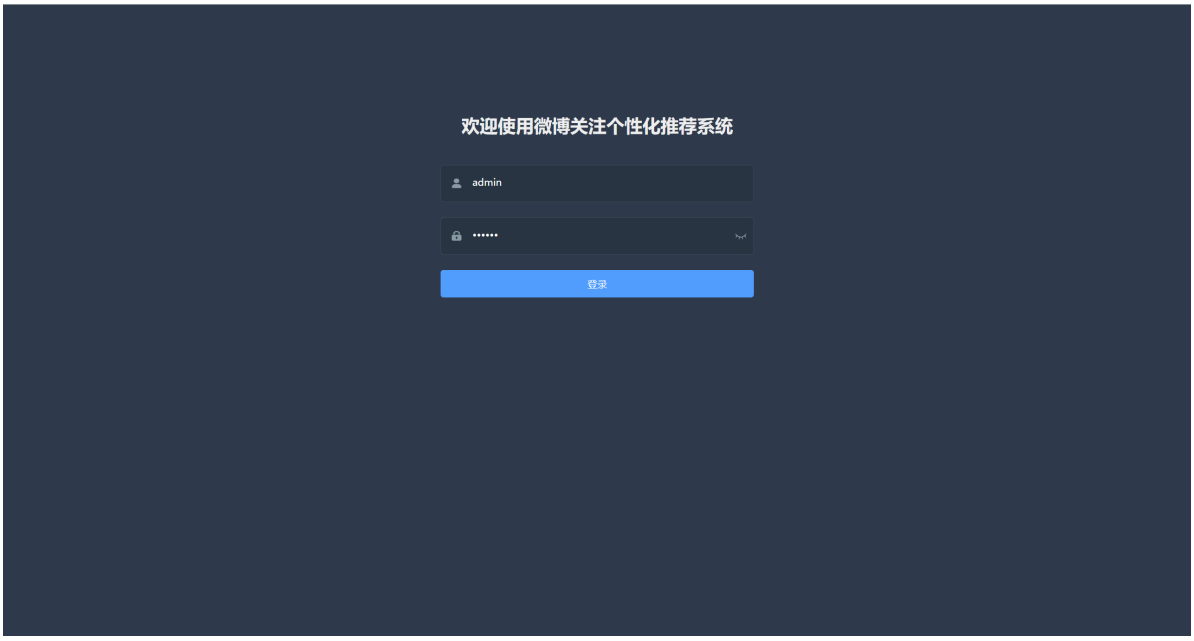
4.1技术路径

系统结构	技术路径
网页前端	Vue
服务端	Flask

系统前端使用vue框架进行设计，采用了vue-admin-template模板提供的脚手架和相关组件来进行制作。

服务端使用python的flask框架进行设计，由于本系统仅作为课堂展示的demo，所以并未使用数据库进行数据存储，而是简单的将微博信息和用户行为信息简单的存储在本地的json文件当中。

4.2登录界面



由于本系统demo并未使用数据库，所以登陆界面仅用一个默认账号进行登陆，登陆界面具备正常登陆界面的一切必备功能，能进行用户名检查和密码检查。

4.3推荐界面

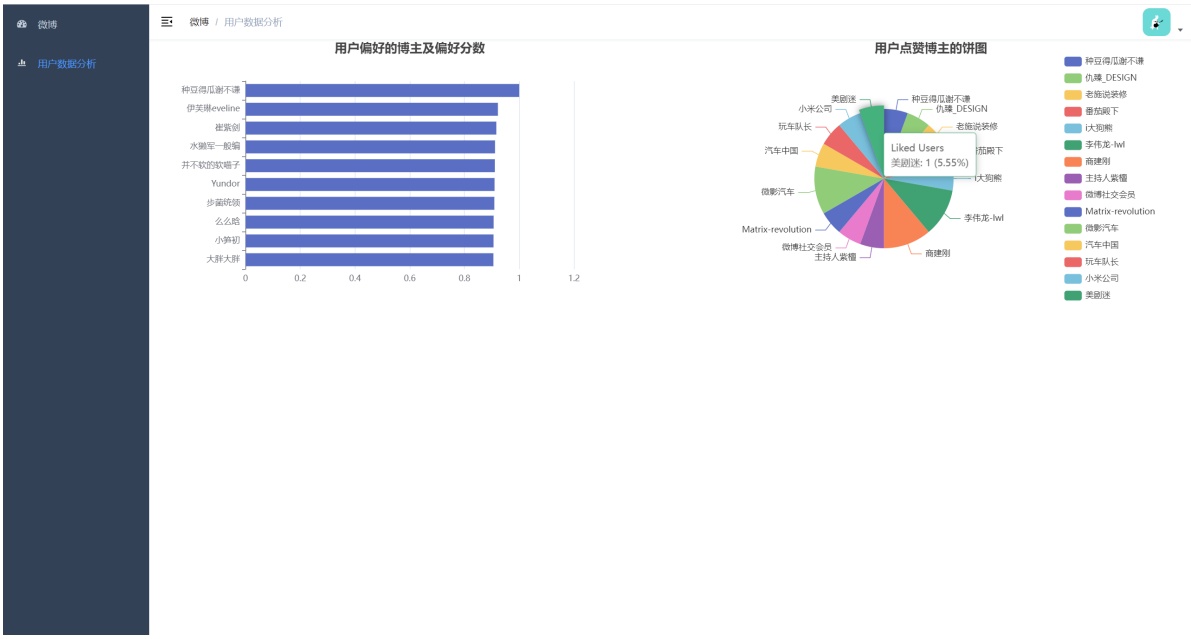
登陆后进入系统的核心界面——微博推荐界面，每次向用户推荐10个微博，推荐算法采用使用gte做embedding的基于内容的推荐算法，每次有5条微博为算法计算出的偏好博主的微博，另外5条为随机挑选的5条微博。

对于推荐的微博，用户可以对其内容进行"点赞"或取消"点赞操作"，每次浏览完一批微博后，点击右下角的刷新按钮，系统就会根据用户的浏览行为（主要是点赞记录）来学习用户的偏好，从而更新用户的行为特征信息，重新构建用户的行为画像，进而实时更新用户的偏好博主，为用户进行下一次推荐。

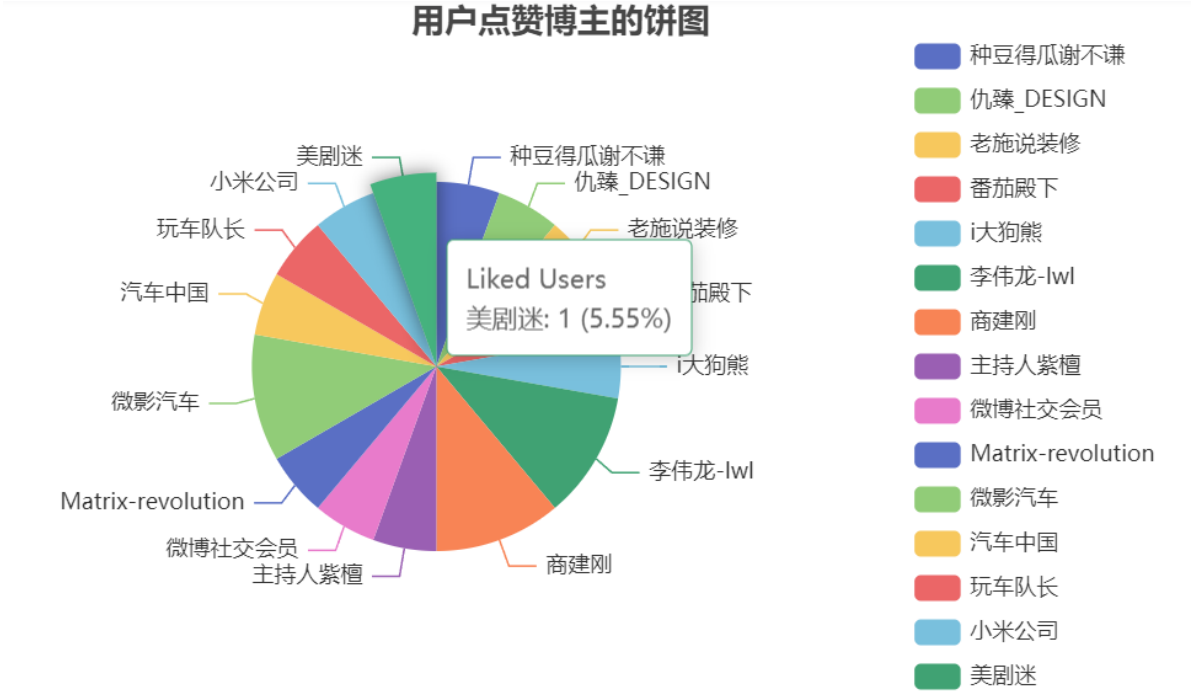
4.4用户数据分析界面

在页面左侧，可以切换到用户数据分析页面，在该界面我们对用户的点赞记录进行可视化展示，以饼图的形式展示在页面右侧。

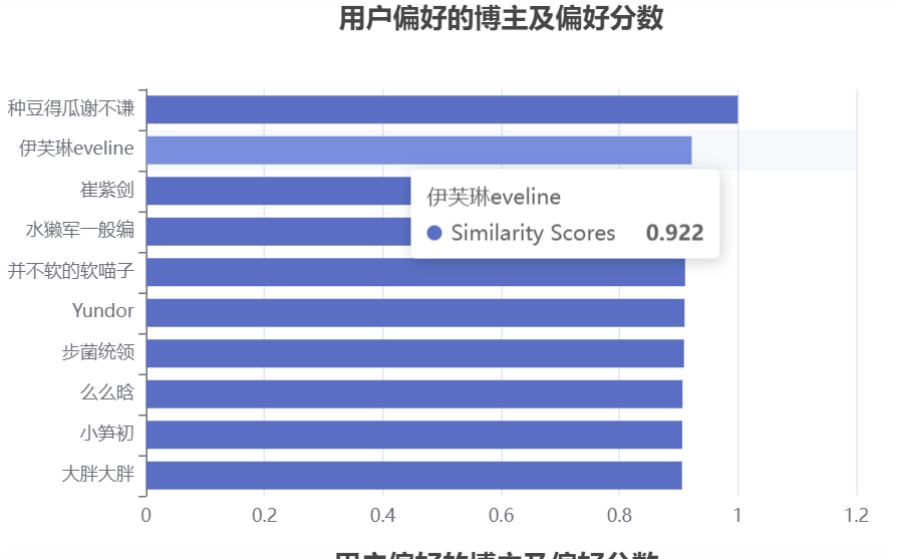
同时我们将基于内容推荐算法根据用户行为特征计算出的用户对各个博主的偏好分数进行排序，并以条形图的形式从高到底展示用户偏好的前十位博主。

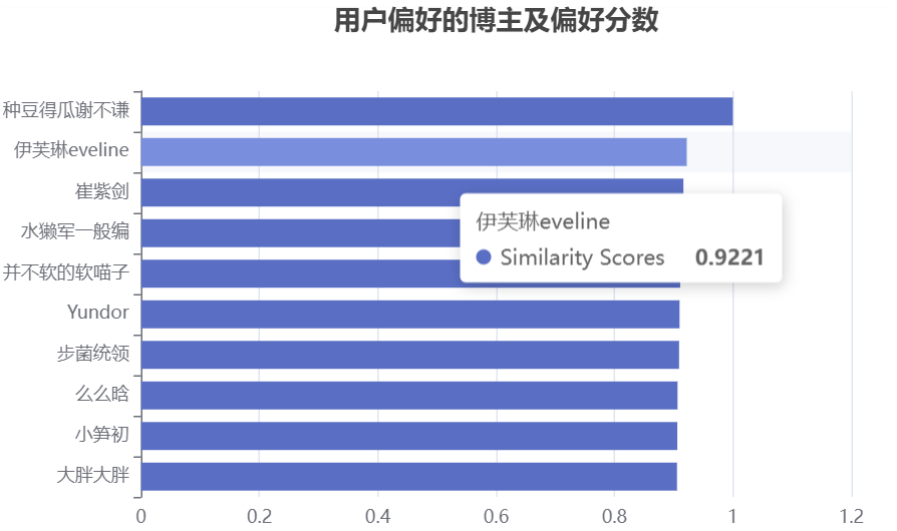


可以看见，对于刚刚我们在浏览时点赞的博主“美剧迷”，用户对其的点赞记录已经被记录在了饼图当中。



用户的偏好博主分数相较于这次点赞之前，也发生了变化（由于先前的记录较多，前十名的偏好博主并未发生改变）。





五、结论

本项目针对微博平台海量信息带来的个性化推荐挑战，通过数据挖掘技术，构建了一个基于用户历史行为、兴趣的个性化微博推荐系统。项目成员分工明确，涵盖了系统设计、数据收集与处理、算法实现及文档编写等多个方面。

我们首先通过微博API成功获取了大量用户推文数据和关注数据，为后续用户行为的理解、内容推荐提供了数据支持。数据预处理阶段，我们处理了日期时间格式、文本清洗和缺失值等常见问题，并针对特定的数据类型如关注者数量进行了特殊处理，确保了数据的质量。并对用户推文数据、关注数据进行了可视化分析。

在模型探索方面，我们使用了协同过滤算法，这是一种基于用户相似性的推荐系统方法。通过计算用户之间的相似度，选择相似的用户作为“邻居”，并利用这些邻居的评分来预测目标用户对未知项目的兴趣。
此外，我们也使用了基于内容的推荐算法。此算法侧重于分析和理解项目、内容本身的特征，以及用户历史行为对这些特征的偏好。推荐系统通过分析内容和用户行为，学习用户的兴趣模型，构建用户画像，并在此基础上向用户推荐内容。通过计算内容之间的相似度并考虑用户的预期兴趣，以提供个性化的内容推荐。

我们设计并实现了基于Vue+Flask的系统，提供了推荐页面，用于展示推荐算法的效果；以及用户数据分析页面，用于展示用户历史行为，并展示用户对各个博主的偏好。

综上所述，本项目针对微博个性化这一现实问题，获取并清洗数据，设计并实现推荐算法，并构建了一个原型系统。未来，我们将继续完善和优化模型，以期实现更高效、更精准的个性化推荐。