基于微博的用户画像

小组成员：周宇田 2120171106
芦美艳 2120171042
康智慧 2120171025
林鹏飞 5720181035

一、 应用背景

当下已经是第三次大革命的时代——“信息革命”，在这个时代里，信息的发展十分迅速，越来越多的人活跃在互联网上。人们可以在互联网上购物，发表言论，获得新闻资讯，寻找娱乐等等。同时在活跃在互联网上的时候，会产生大量的用户信息，浏览信息等等。这些信息数据在一定程度上会暴露个人的隐私，使得个人的信息被不怀好意的人或组织利用进而可能危害个人的心理和生理。但是在另一个方面，如果是在法律允许且不违背道德的情况下，互联网公司可以本着更好的为用户服务的宗旨，不泄露用户个人信息和隐私的同时，合理对用户数据进行挖掘，例如给用户推荐喜欢听的歌等等，尽管现在很对针对性广告投放让用户很头疼。

以上是当今互联网上大信息大数据氛围下的情况，但是在这次研究中，我们不讨论对微博用户的数据进行挖掘是否符合道德标准，是否侵害了用户的权利，仅仅讨论这样一种技术，能够通过对微博用户公开的微博数据的爬取，对这些数据进行挖掘分析，能得出哪些隐藏在数据下面的结论。

二、 用户画像

用户画像就是给用户打上相应的标签，和微博上的微博印象很相似，微博会根据用户平时关注的话题，点赞、评论、转发的内容以及原创内容对用户的标签进行分析。例如，在微博上搜索关键词“邓超”，会有微博印象对邓超的介绍：“孙

俪、变形、臭屁、秀恩爱、过生日、学霸、爸爸、小花"等等，这些就是标签，也就是用户画像。这些标签都是符合邓超身份的，契合他在大众眼中的形象。

我们也会根据抓取到的数据对抓取对象进行微博画像，得出他的标签。同时，我们还会分析出该用户经常活跃的时间段，从而了解用户的作息规律。另外，根据对用户公开微博数据的分析结果，我们也会给出一个对用户的定义，判断用户平时是消极悲观的人还是积极乐观的人。这些都是我们在这个实验中所作的研究。

三、 数据的获取

数据的获取是进行实验研究的关键。好的数据能够让接下来的实验变得顺利，同时也能获得很好的挖掘效果。为此，我们选定的数据属性有以下几种：

time	微博发布时间
app_source	微博发送来源
comment_num	评论数目
content	微博内容
is_repost	是否转发自别人
like_num	被点赞数目
mid	微博 ID
repost_num	被转发数目
rootmid	源微博 ID
rootname	源用户昵称
rootuid	源用户 ID
rooturl	源微博链接

uid	用户 ID，唯一标识用户
uname	用户昵称
url	该条微博的链接

可以看出，以上属性包含了微博所有必要的信息，甚至，如果微博是转发自其他人的微博，也获取了源微博的信息，包括源微博链接等等，在进一步的数据分析中能够提供详细的信息支撑。在这些属性中，最重要的是微博内容和发布时间，通过对大量微博内容的分析，我们能够给用户打上合理的标签，也能判断用户是否是一个积极乐观的人。通过发布时间，我们能够分析用户平时什么时间段活跃在微博上，可以推断用户的日常作息。

四、 情感分析

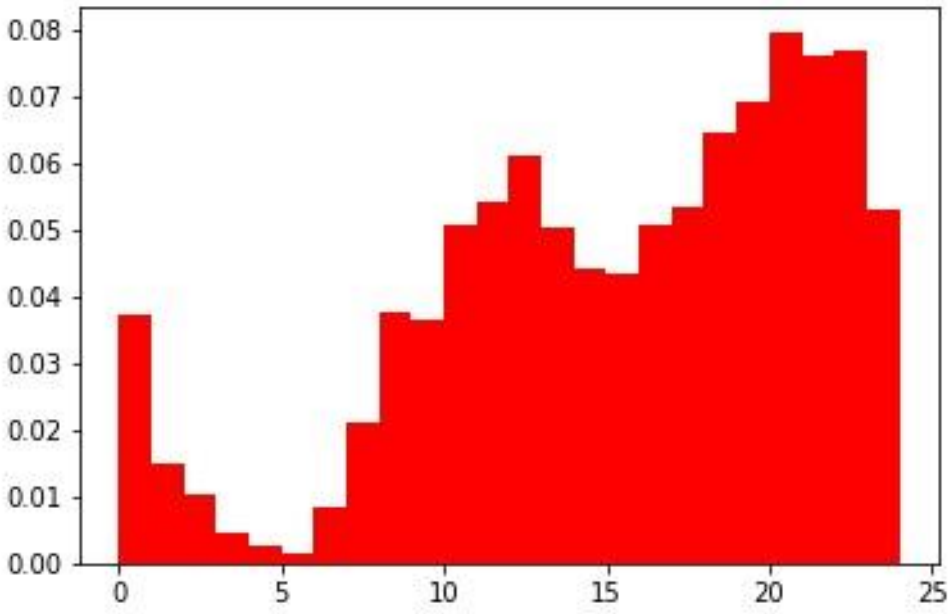
在这个部分，我们采用了 SnowNLP 自然语言处理的 python 类库，可以很方便的处理中文文本内容。在分析过程中，首先对微博用户的每一条微博内容都进行评估，如果通过 SnowNLP 库分析得到的情感值在 0-0.4 范围内，就认为这条微博是消极的，如果在 0.4-0.6 范围内，就认为这条微博内容是中性的，难以判断感情色彩，如果大于 0.6 则认为这条微博的内容是积极乐观的。经过对用户所有的微博处理之后，可以得到用户悲观的微博数目和乐观的微博数目，根据这两项数据来判断用户的情感色彩。

实验结果见 sentiments_all.txt 文件，在文件中，我们可以看到对一些微博号的分析结果。结果显示账号“吴昕”是积极乐观的，账号“快乐大本营”是积极乐观的，账号“思想聚焦”是消极悲观的，“关晓彤”是中性的。

五、 时间分析

在这个部分，针对用户发布微博的时间，我们将一天分为 4 个时段，早上 5 点半到中午 12 点为 morning，中午 12 点到下午 7 点为 afternoon，下午 7 点到晚上 11 点半为 evening，其他时段为 early in the morning。

因此对所有的微博进行统计分析之后得到了对用户上线时间段的分析结果。实验结果见 time_status_all.txt 文件。结果发现很多明星的账号在凌晨依然处于活跃状态，相对而言，普通的微博账号在凌晨仍然在活跃的旧少很多，所有用户在晚上都处于活跃状态。



如上图所示，总体上基本都是在中午和晚上比较活跃凌晨比较不活跃，这是符合人的基本作息规律的。

六、 用户标签

对于这个部分，依旧运用 SnowNLP 对所有微博数据进行分析。调用 WordCloud 方法生成对应的词云标签，对用户进行标签化。结果见词云文件。如词云文件所示，“turbosun”的标签是：孩子、爸爸、一个、今日、妹妹。很符合

孙俪的身份。用户“机器人争霸”的标签是：李晨、机器人、机油、团长、明星、林更新、angelababy。也很符合“机器人争霸”这个节目。所有的这些标签都表明我们成功对用户打上了合理的标签。

项目代码地址：https://github.com/kzh18811708090/kzh