Python 期中大作业报告

薛飞跃 1700017831

一．模块一

1. 爬取微博分为网页端和手机端，相对而言手机端更加容易。
2. 微博的手机端必须添加 cookie 才能爬取更多的微博内容。
3. 微博的网页是动态 json，返回的 json 不能直接用 BeautifulSoup 解析。
4. 微博的评论是瀑布流的形式，评论的第一页的地址为 'https://m.weibo.cn/comments/hotflow?id=%s&mid=%s&max_id_type=0'，其中 id，mid 都可以从微博的动态 json 中解析出来，而第二页开始，网址中再添加一项‘max-id’，并不容易获取。
5. 微博的标签，正文，发布时间，评论数，点赞数，转发量，仔细比对，都存储在网页的 json 中。
6. 为了防止被反爬，需要添加 time.sleep()函数，并且采取随机时间间隔。
7. 微博返回的 request.content 和 request.text 中文都是 Unicode 编码，即使解码后，也出现了大量转义字符，不能很好的解析，暂时没有想出很好的办法。

二．模块二

1. 将爬取的微博的正文用 jieba 分词，用 sklearn 中的 CountVectorizer 进行词频统计，TfidfTransformer 计算 tfidf 值，MultinomialNB 进行文本分类
2. sklearn 中的 Pipeline 可以让将几个过程进行连接，保存模型时，只需要保存 pipe 即可。对测试集也可以直接用 pipe 完成多个过程。
3. 使用停用词，对最终的文本分类预测结果准确性有轻微的提升，但提升不明显。

三．模块三

1. 微博类 p1 实现注册功能时，作为服务端，bind 一个固定端口，监听（listen）并接受（accept）来自客户端的连接，然后 recv 方法接受注册数据（用户的喜好主题和 ip+端口号），解析后保存到本地
2. 微博类 p2 先获取最新的新闻，然后在本地的用户列表中，找到需要推送的用户。此时作为客户端，connect 方法连接用户的 ip 和端口，send 方法发送新闻
3. 用户与上述过程正好相反，注册时，作为客户端，与微博的端口连接，发送自己的兴趣，接受服务时作为客户端，监听并接受来自微博的连接，recv 方法接收消息再解析

四．模块四

1. tkinter 中 Scrollbar 可以设置滚动条，但要注意在特定的组件上安装。此次在 Text 组件中安装滚动条，需要指定 Text 的 yscrollbar 的回调函数为 Scrollbar 的 set，同时需要用 config 函数设置属性，指定 Scrollbar 的 command 的回调函数是 Listbar 的 yview
2. 本次作业中，每次更新，直接加载了本地文件夹，假设了每条微博更新，如模块三中一样，自动推送给用户，用户自动保存在本地。弊处是用户必须时刻接受所有推荐的消息并保存，才可以检查更新。也可以直接调用模块一的爬虫，每次用户点击更新时，主动爬取最新的微博。用户长时间不点击时，不需要接受微博的消息并保存，弊端是，用户若高频率点击更新，即使并未有新的消息给用户，也需要重复爬取微博数据，造成用户刷新时延迟高。