**基于rasa\_nlu的贴吧机器人实现**

**张哲弋**

**第一章 项目背景**

**一、聊天机器人**

聊天机器人(Chatbot)是一种软件应用程序，用于通过文本或文本到语音进行在线聊天对话，而不是与真人代理直接接触。为了令人信服地模拟人类作为会话伙伴的行为方式而设计的chatbot系统通常需要持续的调优和测试，而且许多产品中仍然无法进行充分的会话或通过行业标准图灵测试。术语“ChatterBot”最初是由Michael Mauldin(第一个Verbot的创造者)在1994年创造的，用来描述这些会话程序。

聊天机器人在对话系统中用于各种目的，包括客户服务、请求路由或信息收集。虽然一些聊天机器人应用程序使用广泛的单词分类过程、自然语言处理器和复杂的人工智能，但其他应用程序只是扫描通用关键词，并使用从相关库或数据库中获得的通用短语生成响应。

大多数聊天机器人都是通过网站弹出窗口或虚拟助手在线访问的。它们可以根据使用类别进行分类，包括:商业(通过聊天进行的电子商务)、教育、娱乐、金融、健康、新闻和生产力。

**二、百度贴吧**

贴吧即百度贴吧，是[百度](https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6/6699)旗下独立品牌，全球最大的[中文社区](https://baike.baidu.com/item/%E4%B8%AD%E6%96%87%E7%A4%BE%E5%8C%BA/5752508)。贴吧的创意来自于百度首席执行官[李彦宏](https://baike.baidu.com/item/%E6%9D%8E%E5%BD%A6%E5%AE%8F/125160)：结合[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E/104812)建立一个在线的交流平台，让那些对同一个话题感兴趣的人们聚集在一起，方便地展开交流和互相帮助。贴吧是一种基于关键词的主题交流社区，它与搜索紧密结合，准确把握用户需求，为兴趣而生。

贴吧的使命是让志同道合的人相聚。贴吧的组建依靠搜索引擎关键词，不论是大众话题还是小众话题，都能精准地聚集大批同好网友，展示自我风采，结交知音，搭建别具特色的“兴趣主题”互动平台。贴吧目录涵盖[社会](https://baike.baidu.com/item/%E7%A4%BE%E4%BC%9A/73320)、[地区](https://baike.baidu.com/item/%E5%9C%B0%E5%8C%BA/13841494)、[生活](https://baike.baidu.com/item/%E7%94%9F%E6%B4%BB/18684)、[教育](https://baike.baidu.com/item/%E6%95%99%E8%82%B2/143397)、[娱乐](https://baike.baidu.com/item/%E5%A8%B1%E4%B9%90/279534)[明星](https://baike.baidu.com/item/%E6%98%8E%E6%98%9F/74)、[游戏](https://baike.baidu.com/item/%E6%B8%B8%E6%88%8F/33581)、[体育](https://baike.baidu.com/item/%E4%BD%93%E8%82%B2/223780)、企业等方方面面，是全球最大的中文交流平台，它为人们提供一个表达和交流思想的自由网络空间，并以此汇集志同道合的网友。

随着互联网环境的变化，贴吧的客户端日渐臃肿，并且非超级会员用户在使用贴吧时会收到大量的广告，广告的频率极高，已经对贴吧使用产生了一定影响，图1为相关的两幅贴吧客户端截图。

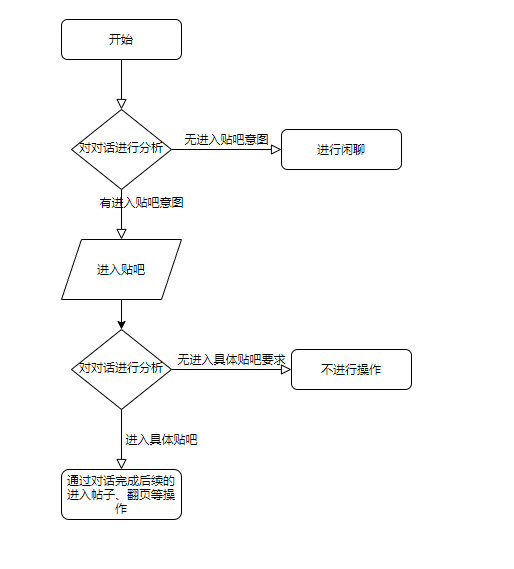
在此背景下，本设计试图通过对话机器人的方式，制作一个迷你的贴吧客户端，并兼具一些其他的对话功能，尝试寻找另一种打开贴吧的方式。

**图****1 当前的贴吧客户端截图**

**第二章 设计与实现**

**一、算法流程**

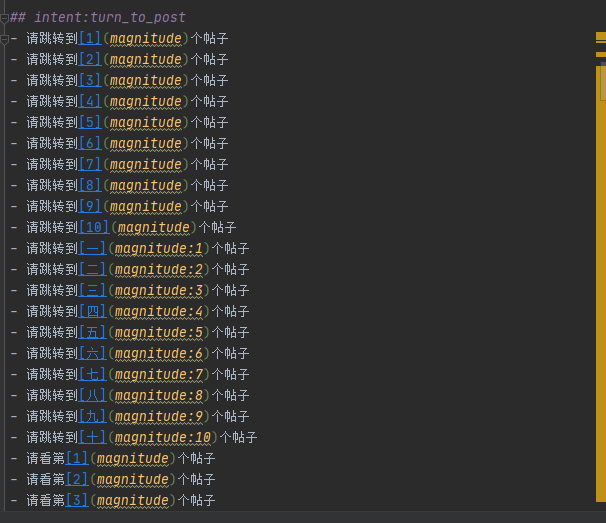
本设计的主要流程为：通过有限自动机的思想，设定一些状态，由此实现从开始、进入贴吧、进入帖子、翻页和返回等操作的实现，并在用户输入无明显意图时不进行状态变化和有效操作。



**图2 项目流程**

**二、数据与框架**

在数据集上，本设计在闲聊对话部分使用了SMP2019(第八届全国社会媒体处理大会)的中文数据集，用于支持一定程度的查询功能的意图识别，例如路线查询、地点查询等。除此之外，本设计所使用的大多数数据集均为程序生成，用于支持贴吧功能的语义理解，例如对于进入帖子功能所要支持的“第一个帖子”“请进入第一个帖子”“我要看倒数第二个帖子”等不同表达的理解，以及对于各种类型的贴吧名称均具有识别能力的要求的转到贴吧对话数据生成，如“我要去抗压背锅吧”“请看魔兽世界吧”等；这些功能的实现均需要大量数据支撑，并且数据本身除了意图外，还需包含entities及value来支持实体检测，需要通过自动化生成的方式获得。



**图3 部分数据集；左为部分SMP2019数据集，右为部分自生成数据集**

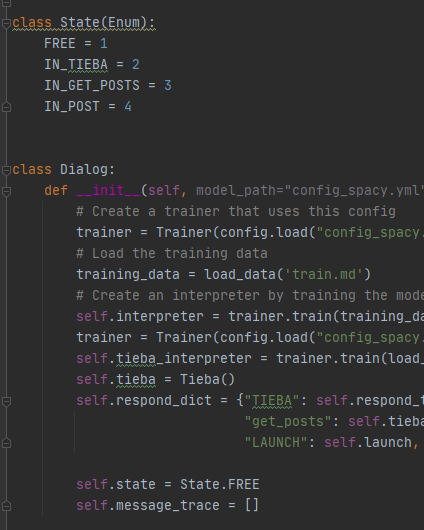
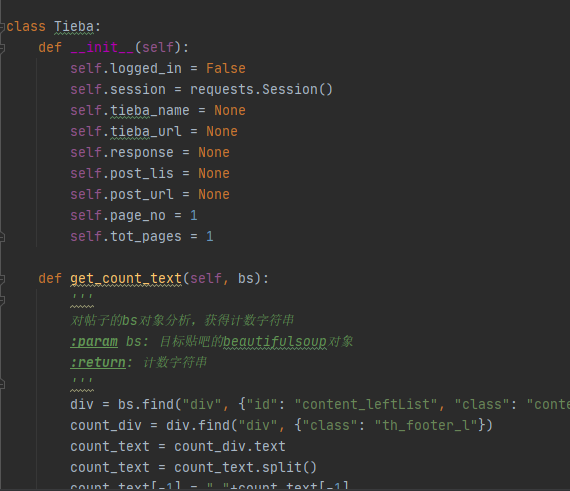
在框架上，本设计在python上完成，主要基于rasa\_nlu框架实现，除此之外由于爬虫、正则匹配、聊天软件api调用等相关需求使用了许多其他第三方库，表1为本设计所涉及到的部分第三方库版本（详见requirements.txt）

|  |  |
| --- | --- |
| 库 | 版本 |
| cn2an | 0.5.5 |
| telegram | 0.0.1 |
| requests | 2.24.0 |
| rasa\_nlu | 0.15.1 |
| beautifulsoup4 | 4.9.1 |

**表1 用到的第三方库版本示例**

**三、具体实现**

在实现上，除了用于临时生成数据集的代码、用于整体流程实现的main文件等，本设计主要通过实现以下的类，对相关功能进行封装来达成目标：实现贴吧访问的Tieba类、控制对话的Dialog类，及控制状态的State枚举类。



**图4 主要功能代码片段**

在实现中，Tieba类主要依靠爬虫技术实现，通过对于贴吧网页端的HTML分析，得到构建迷你贴吧客户端所需要的部分实现；Dialog类则与State枚举类密切相关，通过维护一个由（状态，输入)到（状态，操作函数）的字典，利用有限自动机的思想进行实现；在具体的对话机器人平台上，由于微信平台的itchat、QQ平台的CoolQ等相关api已被禁用（其中CoolQ在一月内被禁用），本设计采用了Telegram API进行实现。

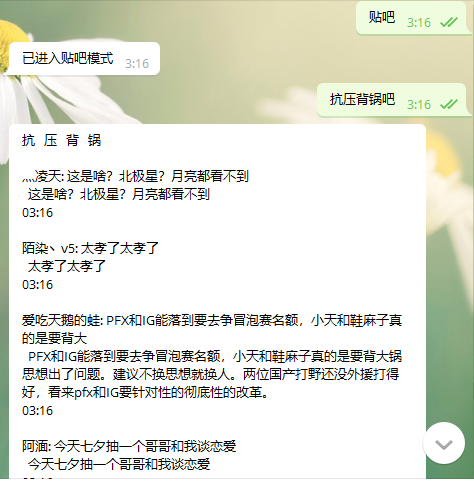
**第三章 功能展示**

经过测试，本设计所实现的对话机器人首先可以进行一定程度的基础查询操作，包括地点查询、菜谱查询、路线查询等：

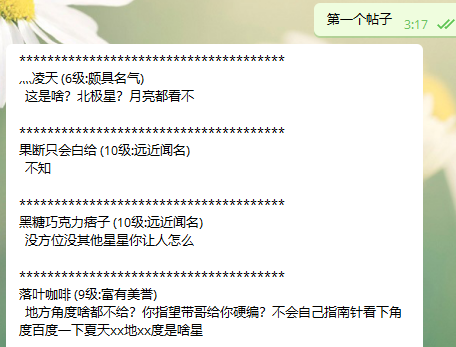


**图5 基本查询功能**

在主要功能的测试上，本设计所实现的贴吧功能可有效支持贴吧的阅读操作，包括进入贴吧、进入帖子、在帖子内进行翻页等，并通过维护操作栈，支持返回上一步的操作：



**图6 贴吧浏览功能**



**图7 帖子浏览功能**



**图8 返回功能**

**第四章 总结与不足**

在本设计中，我实现了具有一定基础功能和贴吧阅读功能的对话机器人，在一定程度上对贴吧客户端日渐臃肿的问题进行了尝试解决；但由于问题及百度登录系统复杂性限制等原因，本设计并未实现预计的通过requests的session进行账号登录管理，实现登录甚至多开小号及发帖回复等功能，这也是本设计的一个发展方向。