**智能聊天机器人**

**远程科研项目**

课程名称：金融类智能聊天机器人

专业：计算机技术与科学

学生名称：李哲龙

日期：2019年8月17日

**目录**

**摘要**

1. **任务背景**
   1. **聊天机器人**
   2. **支持向量机**
2. **任务分析**
3. **项目设计和代码实现**

**3.1 自动闲聊回复**

**3.2 意图识别和实体抽取**

**3.3 意图和模糊词的训练集制作及训练**

**3.4 状态机和多轮多次问询**

1. **成果展示**
2. **总结**

**摘要**

随着自然语言处理与机器学习技术的发展，两者结合所诞生的各种产品，使得即使是对相关技术了解不深的人们也可以更加便捷精准的通过日常对话的方式，来获取用户个体所需求的信息，节省了用户的繁琐操作与时间成本。本项目的设计初衷即为此，让用户得以通过最贴近生活，他们最熟悉的方式来达成查询信息的目的。

本项目主要的应用背景为金融领域。在用户试图查询某支股票的某项信息时，他们通常需要通过复杂繁琐的查询步骤和专业软件工具来进行查询，增加了新手的入门难度。通过将前台的繁琐操作转入后台，并配合回复机器人的自然语言问询，可以极大地降低查询难度进而照顾到更多人的使用体验。

项目的编程语言为Python，使用的核心库为Spacy，Rasa\_nlu，iexfinance，等。实现的核心技术为：基于正则表达式的格式化回复；通过关键词，正则表达式和Rasa\_nlu多种技术结合的意图识别与模糊词合并；基于Spacy的实体类抽取；基于状态机和待定任务转换的多轮多次查询技术；基于iexfinance api的联网股票实时查询。后续还将集成到telegram上，使其成为一个真正意义上可与用户自动互动的聊天机器人。

除去全程可以进行与机器人的闲聊外，在用户登录后，用户可以根据自己的需求通过先后声明股票名称，和查询内容的方式达成相应的查询目的。

**关键词：Spacy, Rasa\_nlu, telegram, 支持向量机，状态机，待执行项目，多轮多次问询，正则表达式，意图识别，实体类抽取**

1. **任务背景**
   1. 聊天机器人

聊天机器人（Chatterbot）是经由对话或文字进行交谈的计算机程序。能够模拟人类对话，通过图灵测试。聊天机器人可用于实用的目的，如客户服务或资讯获取。有些聊天机器人会搭载自然语言处理系统，但大多简单的系统只会撷取输入的关键字，再从数据库中找寻最合适的应答句。

目前，聊天机器人是虚拟助理（如 Google 智能助理）的一部分，可以与许多组织的应用程序，网站以及即时消息平台（Facebook Messenger）连接。非助理应用程序包括娱乐目的的聊天室，研究和特定产品促销，社交机器人。

聊天机器人（chatterbot）是一个用来模拟人类对话或聊天的程序。“Eliza” 和 “Parry”是最早期的非常著名的聊天机器人。它试图建立这样的程序：至少暂时性地让一个真正的人类认为他们正在和另一个人聊天。

聊天机器人产生的原因是，研发者把自己感兴趣的回答放到数据库中，当一个问题被抛给聊天机器人时，它通过算法，从数据库中找到最贴切的答案，回复 给它的聊伴。

此外，一些较为成功的聊天机器人的成功之处在于，研发者将大量网络流行的俏皮语言加入词库，当你发送的词组和句子被词库识别后，程序将通过算法把预先设定好的回答回复给你。而词库的丰富程度、回复的速度，是一个聊天机器人能不能得到大众喜欢的重要因素。千篇一律的回答不能得到大众青睐，中规中矩的话语也不会引起人们共鸣。

聊天机器人应用场景丰富，在通讯平台，教育，在线服务等领域已被广泛使用

* 1. 支持向量机

支持向量机（Support Vector Machine）是一种监督学习（supervised learning）的二元分类（binary classification）的广义线性分类器（generalized linear classifier），其决策边界（margin）是对学习样本求解的最大边距超平面。

SVM使用铰链损失函数（hinge loss）计算经验风险（empirical risk）并在求解系统中加入了正则化项以优化结构风险（structural risk），是一个具有稀疏性和稳健性的分类器 [2]  。SVM可以通过[核方法](https://baike.baidu.com/item/%E6%A0%B8%E6%96%B9%E6%B3%95/1683712)（kernel method）进行非线性分类，是常见的核学习（kernel learning）方法之一。

在本项目中，Rasa\_nlu会根据训练样本，使用SVM来训练模型，最终识别用户输入自然语言的具体意图。

1. **任务分析**

本项目的目的是搭建一个可按用户输入信息进行金融信息查询的聊天机器人，同时尽可能保证模型的鲁棒性（Robustness）。

根据几周的课程学习，以及老师的知识梳理和讲解，本项目可分为以下几个部分：

1. 对于各种数据进行预处理，训练模型
2. 对于用户输入信息进行意图识别，判断当前自然语言的意图
3. 根据意图给予适当的回复，并转换相应的状态阶段
4. 使用提取出的有效信息进行联网查询并同时反馈

在具体实施过程中，为了保证随时可以与对话机器人闲聊，我们设置闲聊模块的优先级高于问询流程，即意图识别后首先判断是否属于闲聊信息，在不符合闲聊预先设置的规则后再继续之前问询的流程。同时，错误的输入格式不能中断之前的问询流程以及实现存储的信息。

1. **项目设计和代码实现**
   1. 自动闲聊回复

由于本项目中的闲聊功能不作为主体功能，所以设计较为简单直接，运用了正则表达式来匹配用户输入信息中的关键词，之后使用预先配置的固定回复句式进行随机回复。

具体可分为两步。第一步判断输入信息是否符合预先设计的固定句式：

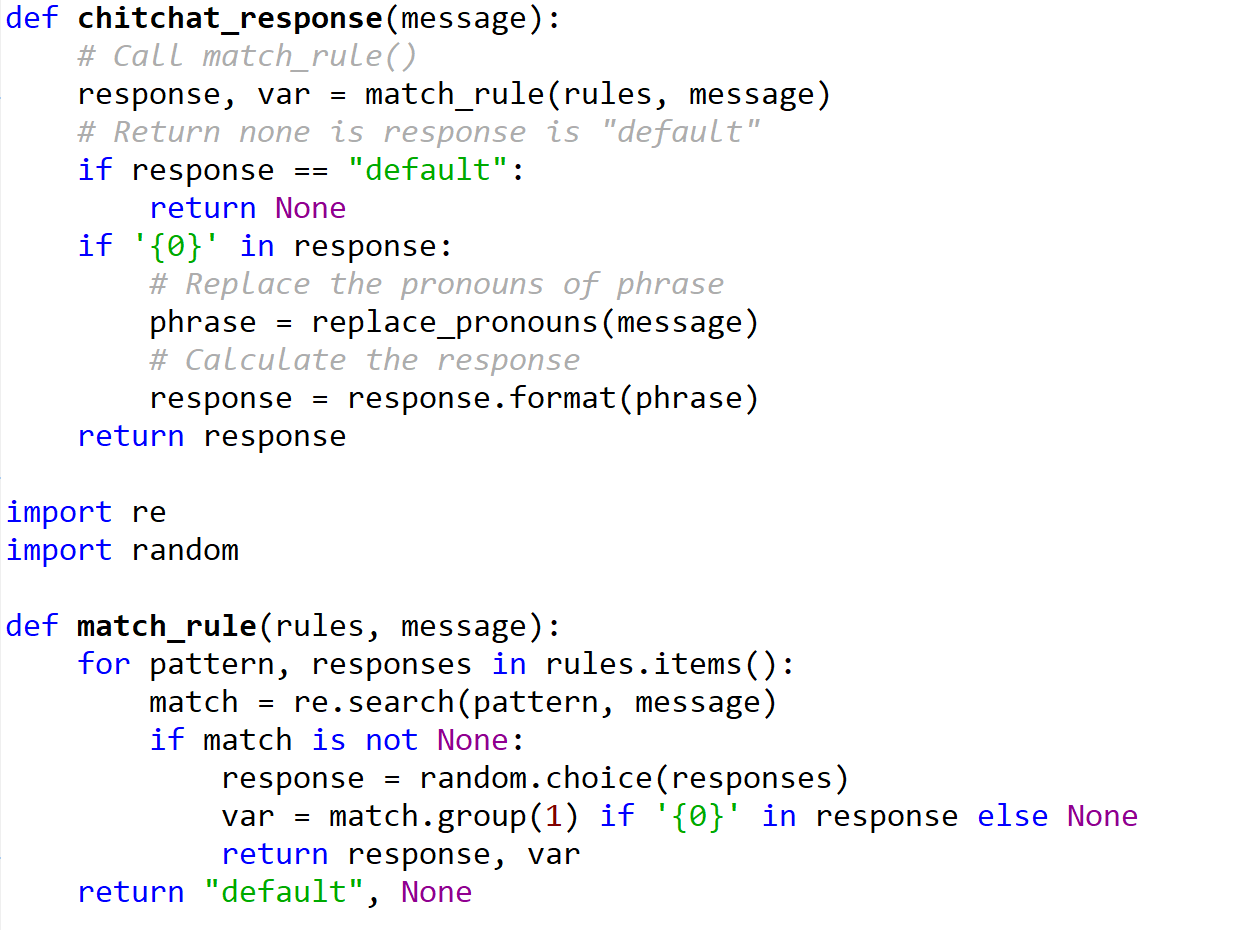


图1 闲聊中的固定句式匹配

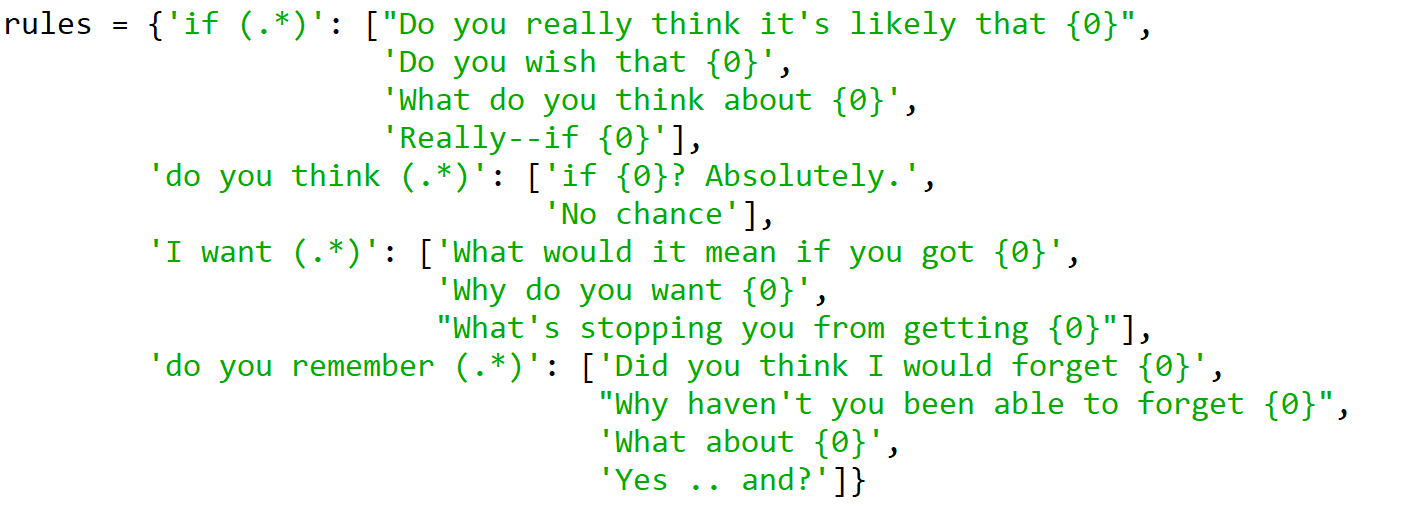


图2 预先设置的固定句式

第二步根据所使用的自然语言语法特点置换人称：

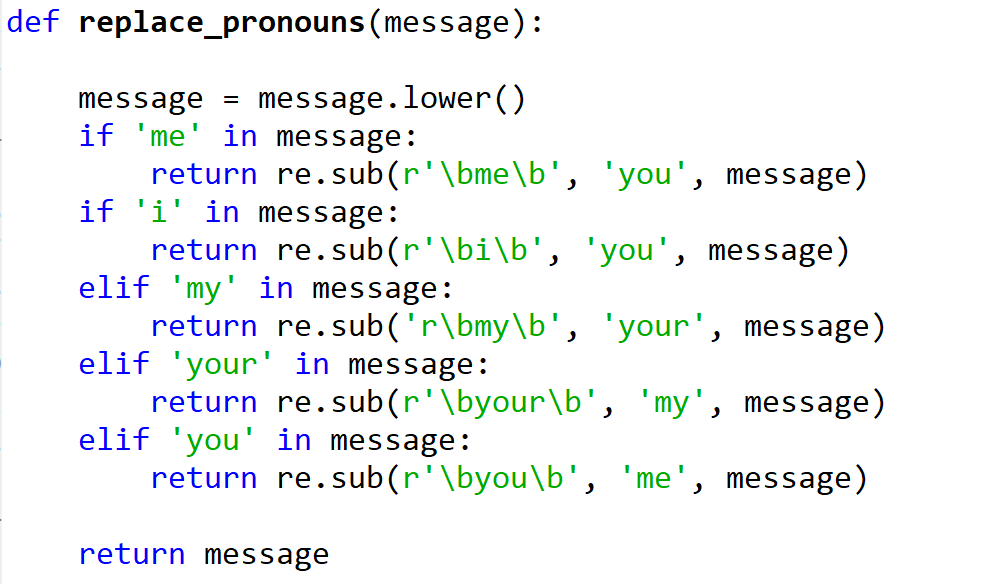


图3 初步的人称替换

此处的人称置换方式较为简单，可能存在漏洞。然而此处的功能仅作为闲聊功能的样例，其具体实现方式并不完善，在此不做过多赘述。

* 1. 意图识别和实体抽取

在每次接收到用户的输入信息之后，首先对信息进行意图的判别：

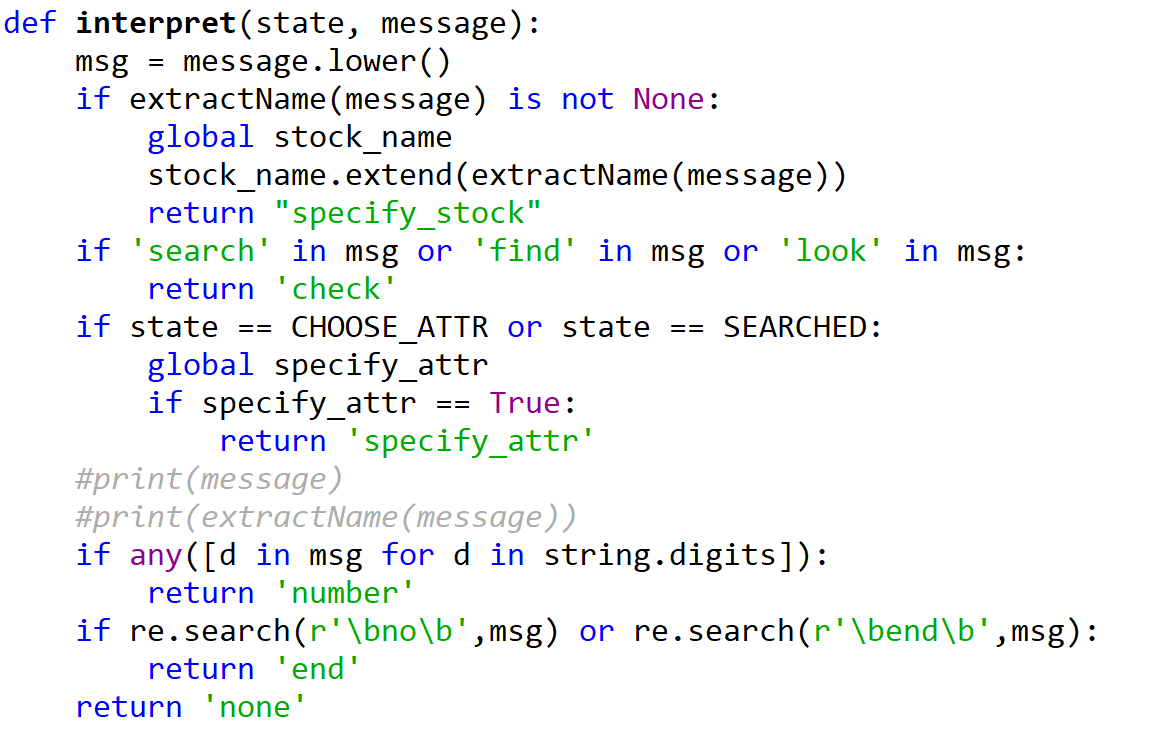
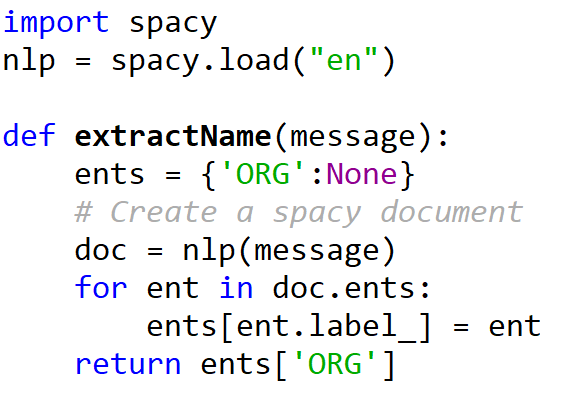


图4 结合状态的意图识别

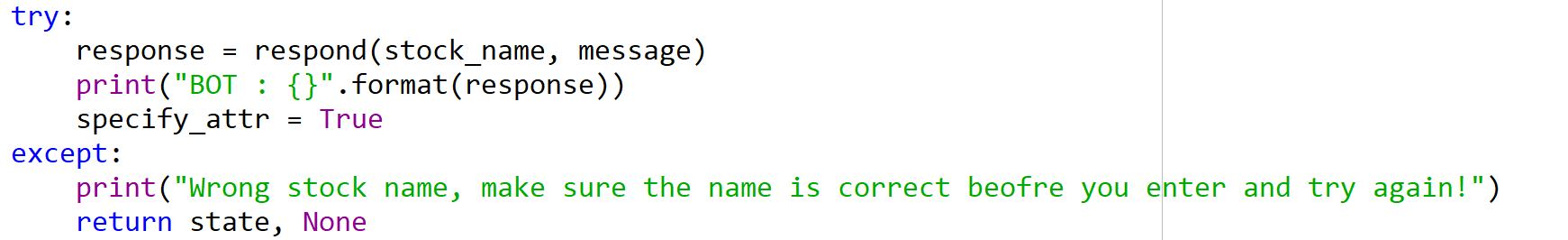
其中，识别意图的方式有三种：

第一种最为直接，根据输入信息中是否含有特定的字段。

第二种运用到了实体类抽取，即图中的extractName函数，运用Spacy解析句子结构后，提取出句子中的ORG主体类。在本项目中，进入申明股票名称阶段时将会用到，并同时存储/更新股票名称。

图5 实体提取

第三种则在send\_message主函数中首先提取用户声明的想要查询的股票的某种特征信息名称(attr)后，尝试通过iexfinance联网获取该种股票对应的特征信息，若成功则判定声明的特征名称符合规则，失败则请用户重新声明：



在不手动输入与iexfinance api所提供的各种key名称时，本项目的选择为在保证该查询可以成功得出结果的一种合理选择。

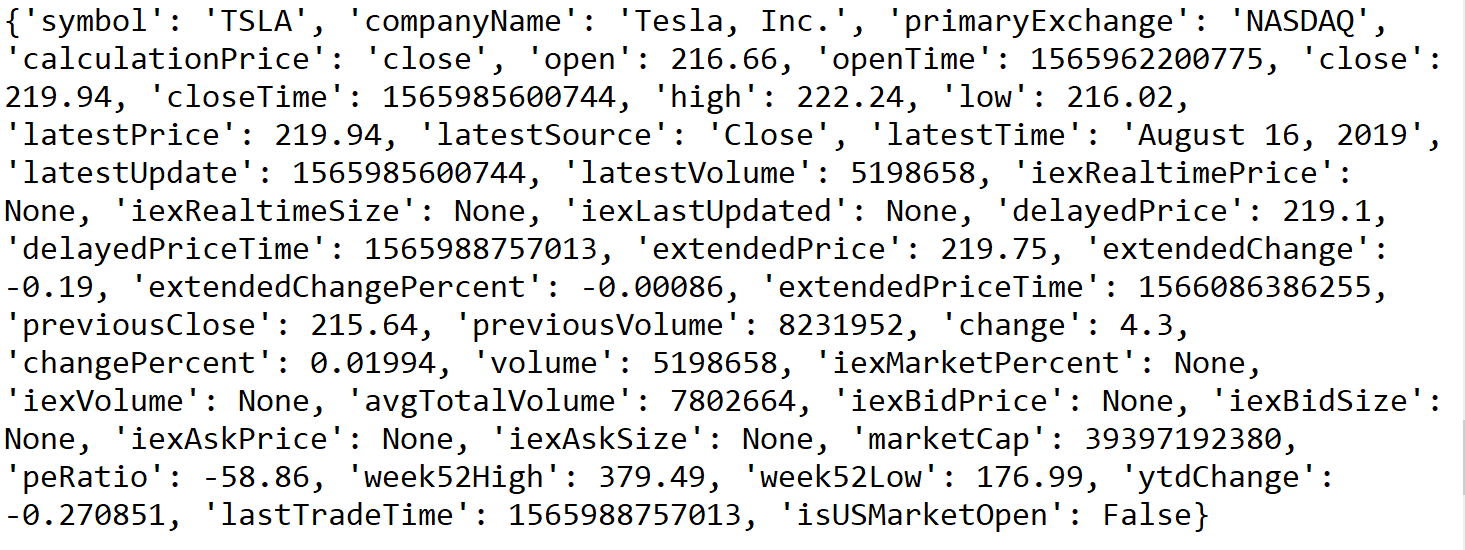


图6 运行api get\_quote() 的查询结果

可以想象的一种更为简单，但相对更硬性编码的选择是首先完成对已知股票的查询并存储，再比对新问询的特征信息名称（attr）是否为存在于预先存储的字典中。

在提取输入信息的特征信息名称（attr）环节，我用到了Rasa\_nlu以及实现配置的训练集demo-stock来训练模型，并用模型分析信息句子的意图。这些内容将在下一个小节中讲解。

* 1. 意图和模糊词的训练集制作及训练

模糊词的训练旨在提升模型的鲁棒性，即在用户输入关键信息出现错别字的时候依然可以准确的揣摩出意图并给予准确的回复：

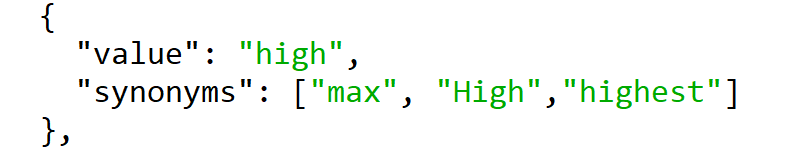


图7 训练集中的模糊词示例

意图识别的训练在本项目中用于提取句子中用户想要了解的指定信息（attr）名称：

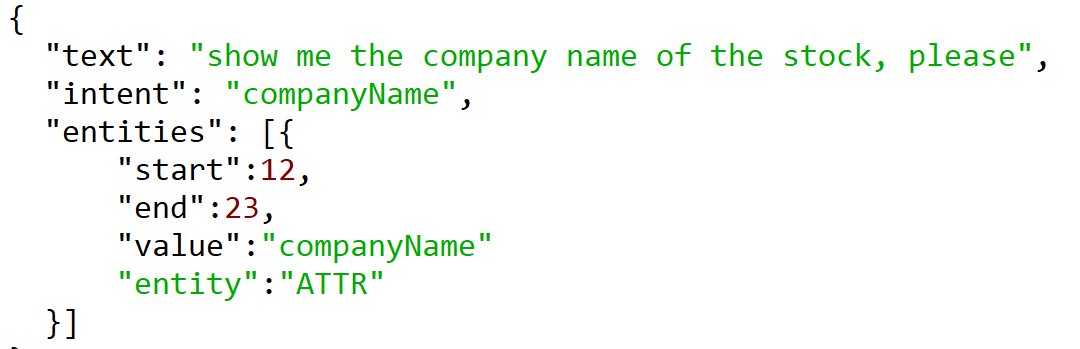


图8 训练集中的意图识别示例

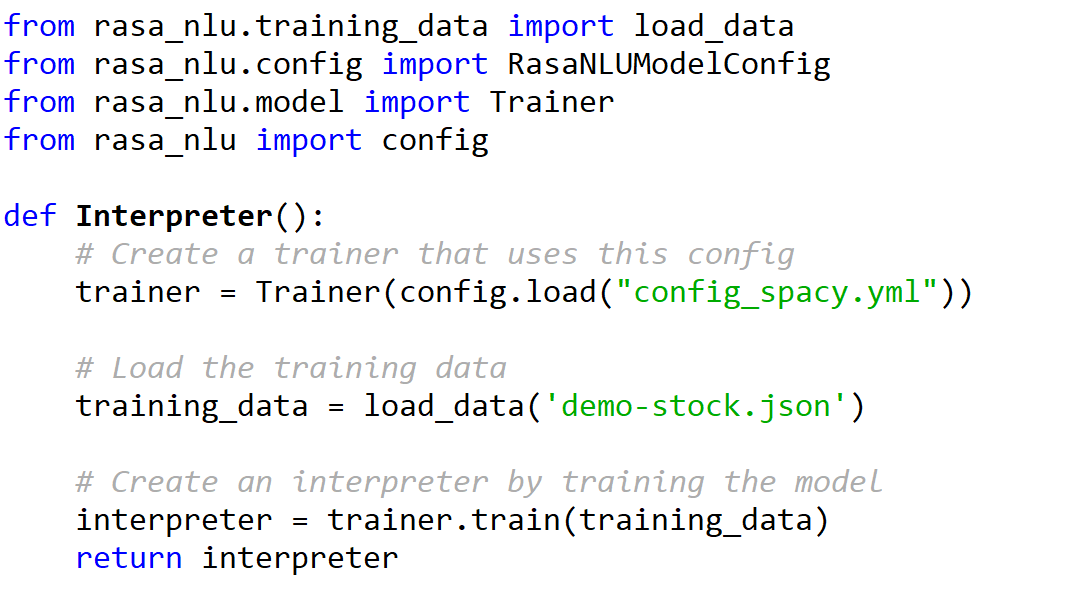
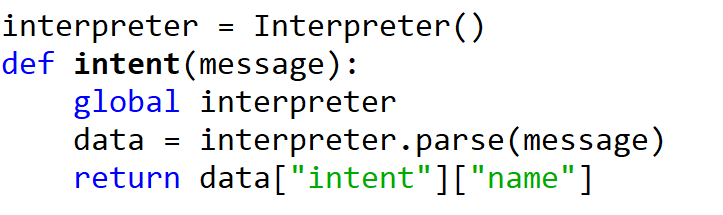
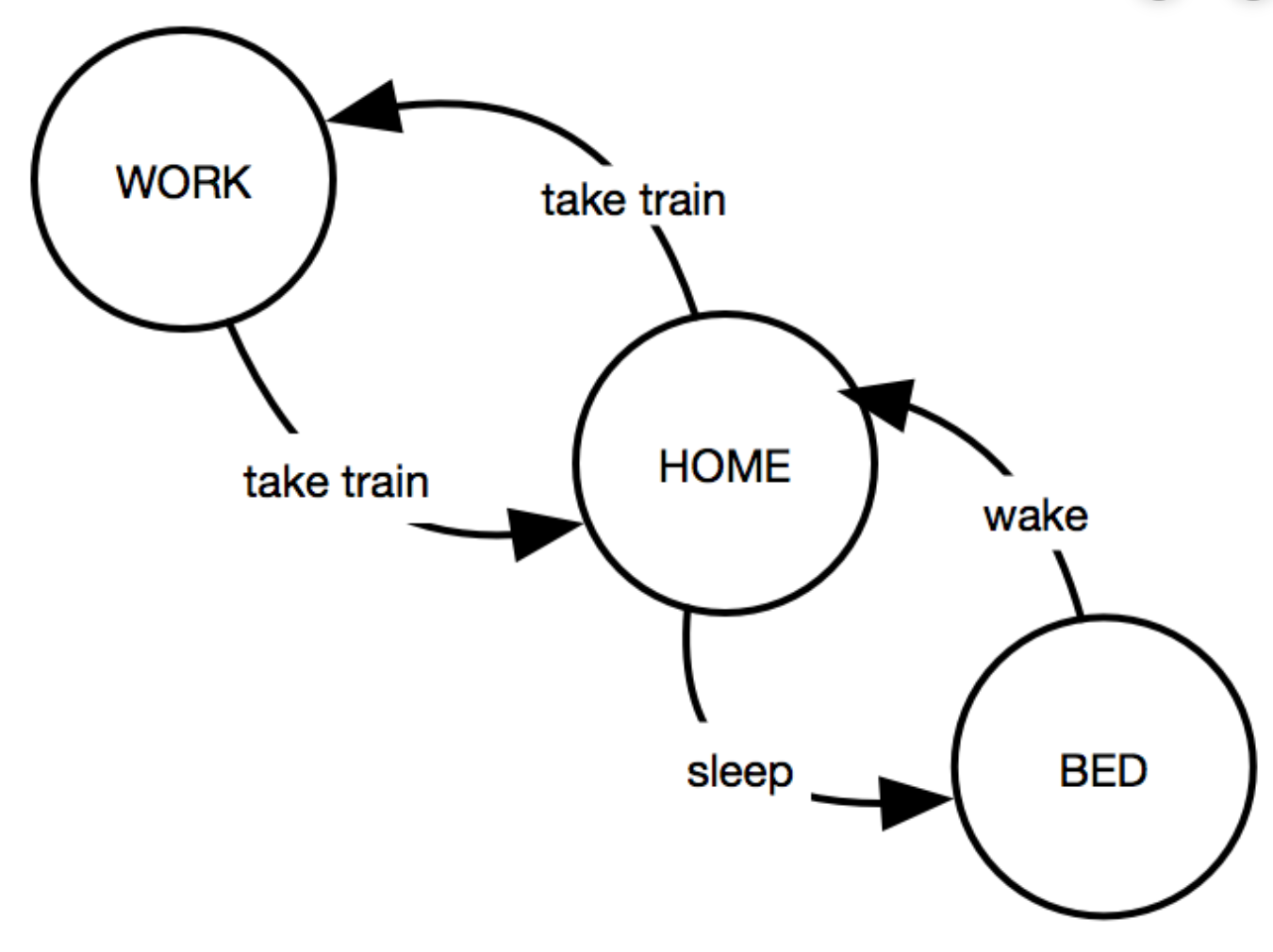


图7 利用Rasa\_nlu进行模型训练

图8 提取意图名称

* 1. 状态机和多轮多次问询

为了达到多次问询的消息相互之间有联系，问询有连贯性，本项目采用状态机来处理。状态机（state machine）的特点是对于每一个状态有一个定义，在达成不同的条件时可以从当前状态转换到由条件（condition）决定的下一个状态。其中，上文介绍的意图识别功能可以很好的充当区分条件，在此基础上本项目还实现了待执行项目的处理功能，以及随时中断当前问询，开启另一支股票的查询功能。

图9 状态机与条件示例

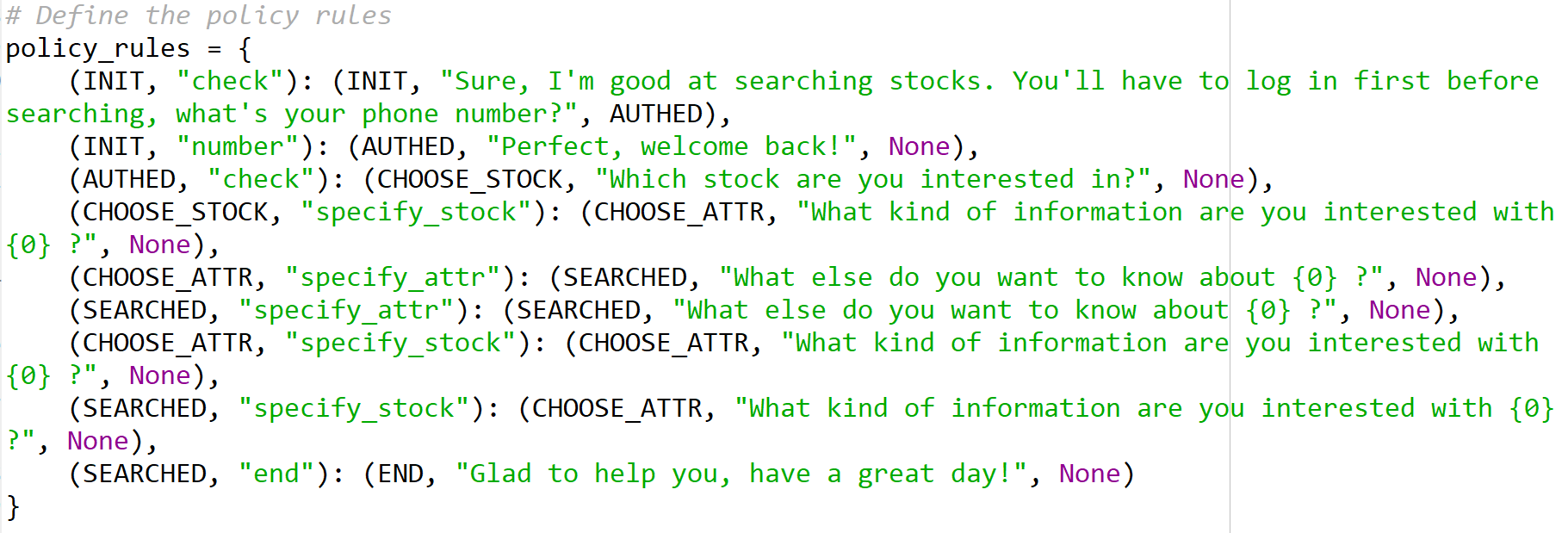


图10 包括待执行项目的状态机示例

首先在符合问询标准之前，出于对安全性的考虑和控制请求接口的QPS，对于任何信息查询，都需要经过登录验证，在此做简单示范，即需要用户提供有效的电话号码，否则将无法进行到下一个状态。

待定执行任务的设置，使得状态机实现“认证”功能，显然这只是可以实现功能的冰山一角，还有许多其他有益的功能可以通过处理待执行项目来实现。

1. **成果展示**

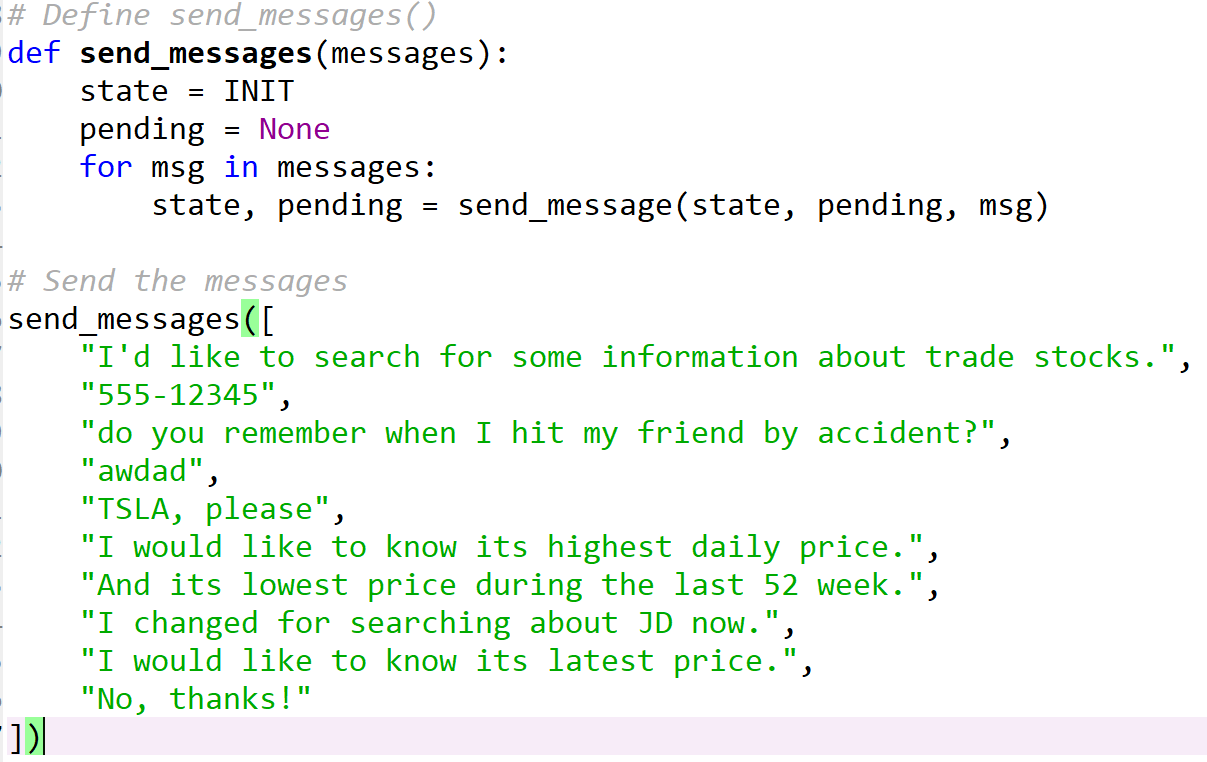


图11 模拟用户的多轮对话

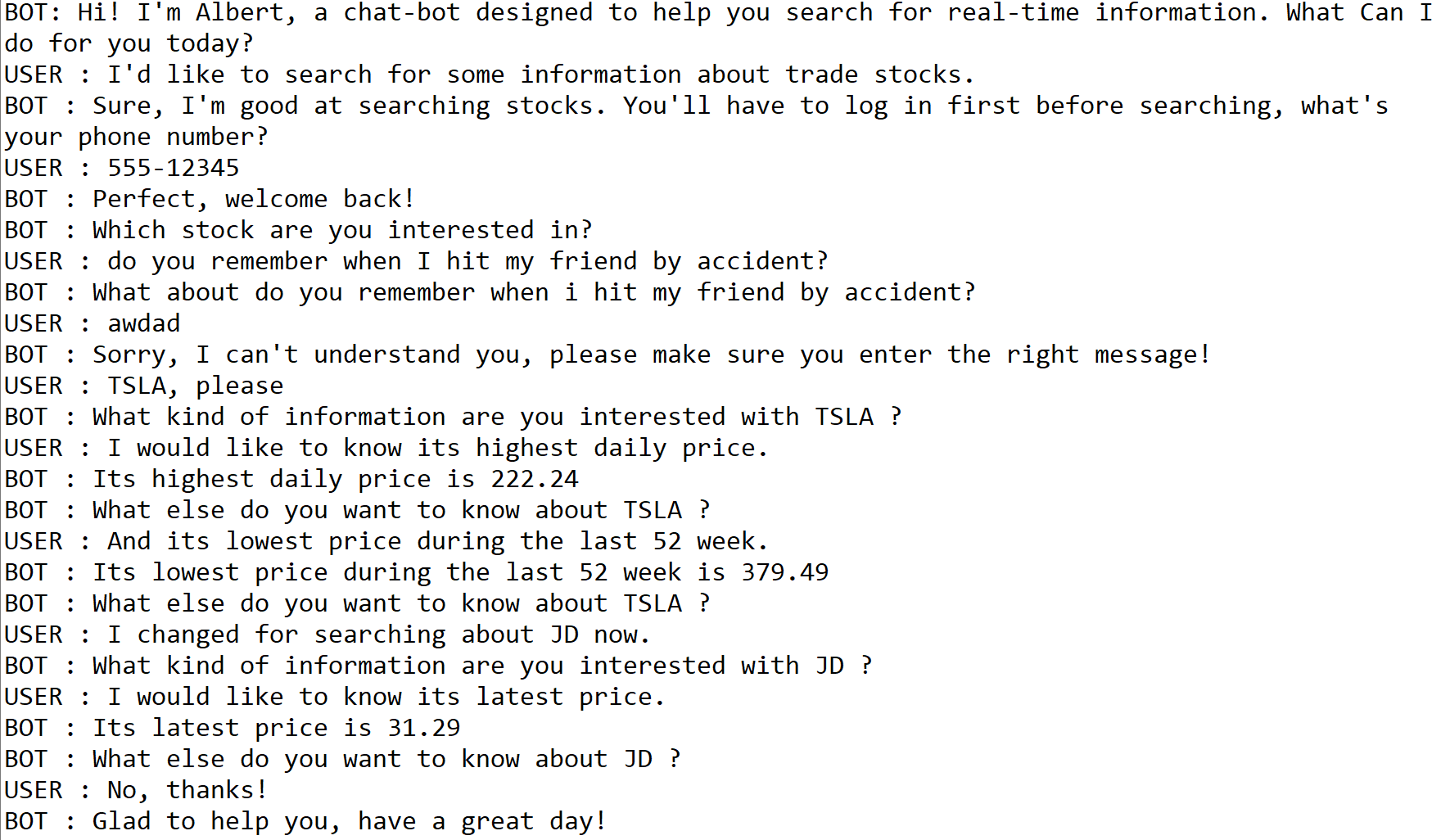


图12 运行结果

上图展示了前文所描述的诸多功能，包括意图识别，实体类抽取，闲聊，中断当前查询，改为查询另一支股票，通过iexfinance api联网查询股票相关信息。其中，如果输入信息的意图不包含在意图识别函数或闲聊中，或者意图正确，输入信息不够准确，都会有相应的报错语句，并且不会扰乱整个问询过程，使用者只需要根据提示继续先前的进展。

**总结**

本次“聊天机器人”的远程科研项目为期一个月，指导老师张老师由宏入微，由浅及深的介绍了人工智能AI的发展历史，对近代科研成果做了概述和举例，同时讲解了自然语言处理NLP的研究背景，核心技术，应用场景等，完善了我对两者的认知。

除去项目中使用的技巧，我还学会了如何通过实体识别和意图识别的Python库Spacy和Rasa\_nlu，结合适当的函数达成各种功能，包括甄别用户语言中的否定实体；利用Sqlite3调用数据库并筛选出自然语言中所需信息；熟悉了Python中正则表达式的应用等等。

本项目很好的将机器学习，自然语言处理结合，让我在动手写代码的过程中加深对两者技术理念和实现方式的认识。编程过程中出现的各种Bug也促使我熟悉了python许多基础元素如字典的特别之处与特定用法。

有些遗憾的是原定将集成到微信的计划因为受到微信出于安全考虑禁用网络版的阻拦而搁浅，而集成到Telegram的想法也因为网络原因暂时无法实施，后续回到美国会跟进。