人工智能远程研究总结报告

赵周杰

目录

一、研究背景和介绍

人工智能成为人类研究的目标有很多年了。自 1956 年这个词语被创造以来，一代代的科学家将其作为奋斗目标。从最开始的“一个夏季的研究”，二十世纪七十年代的“冬天”，到现在大大小小的无数细分研究项目，人类已经认识到人工智能的复杂。同样的，伴随着技术的进步，虽然科幻小说中那样智慧化身班的机器仍然遥不可及，但在很多限定的领域，人工智能已经有了超出人类的表现。

人工智能已经发展出了相当多的算法。如果说 K Nearest Neighbourhood（KNN） 和 Support Vector Machine 还有原始的 Classifier Tree 还是数学在人工智能领域的延伸， 从 Adaboost 和 Random Forest 开始，人工智能已经学会学习，不断地改善自己。而 Convolutional Neural Network(CNN)神经网络更是引领着人工智能飞速的发展。配合着大数据时代的到来，海量的数据使得神经网络的训练变的越来越容易，residual network 也让更多层数的深层神经网络的实现成为了可能，人工智能的准确率变的越来越高，也让越来越多的项目可以应用到生活中。

当一样科技从研究走入生活，它的需求就猛然增加了。从各大高校、公司对人工智能领域人才的待遇就可以知道这项技术所带来的巨大的价值。在生活的方方面面，人工智能在人们不经意间已融入其中。像是热点推送，智能识图，语音识别，家居小精灵，人工智能的普及其实早就超出了一般人的想象。

这次的研究项目，主题就是人工智能当前一个热门的分支，聊天机器人。人类的语言中夹杂着太多的信息，即使人与人之间的对话也常常发生误解，存在歧义。这些对于机器来说更是一个巨大的挑战。如何识别一句话中的实体，又如何判断语义？在这之后又如何进行回答？即使是 IBM 的 Watson 也还不能完美的做好这些。但当我们将范围缩小到一个特定的领域，即使是初学者也能做出像模像样的聊天机器人。这便是这次研究项目的目的。

二、相关技术和方法

I. 正则表达式匹配

运用正则表达式来匹配字符串，用来寻找一定格式的子串，提取关键词，比如说电话号码，邮箱。正则表达式在一定程度上耗时相当少，所以将其与人工智能技术结合可以提高效率，减少耗时。

```
# match phone number
match = re.search('(\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)\s*\d{3}[-\.\s]??\d{4}|\d{3}[-\.\s]??\d{4})',
                  message)
```

电话号码的匹配

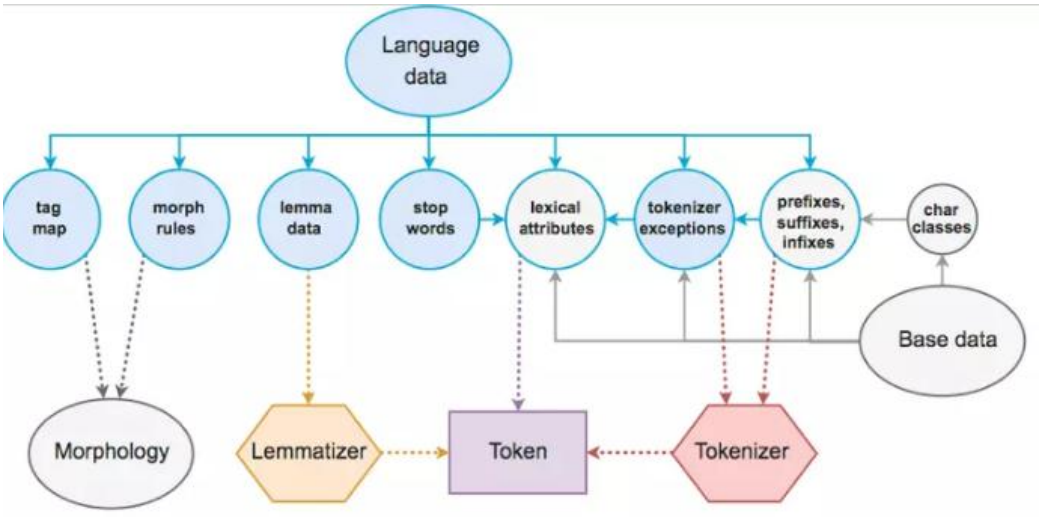
II. 最近邻分类法、词向量和支持向量机

运用最近邻分类法和支持向量机配合上词向量可以将输入语句和之前的训练数据进行对比，从而获取输入语句的意图，获取实体，依赖分析。

III. 状态机

状态机可以控制机器人的状态信息，从而实现层层递进，循环或者待定等不同的结构，让多轮多次查询变的可能，这也是让机器人变的更加人性化，增加复杂度。

以上技术部分将依赖于 spacy，rasa_nlu 这两个 package 的支持。



我这次做的事关于股票信息查询的聊天机器人，关于股票的信息将依赖于 iexfinance 的 API。

```
In [1]: from iexfinance.stocks import Stock

In [2]: tsla = Stock('TSLA')

In [3]: tsla.get_price()
Out[3]: 343.92
```

```
In [4]: batch = Stock(["TSLA", "AAPL"])

In [5]: batch.get_price()
Out[5]: {'TSLA': 343.92, 'AAPL': 174.24}
```

关于 python 与微信的对接，应用了 wxpy。

首先，新建一个训练数据集，用 rasa_nlu 进行训练来获取意图。运用状态机来实现多轮查询时需要的多轮信息。当用户所给的信息或者意图不是查询所需要的时候，进入 chitchat 对话并保持状态和等待状态。Chitchat 运用了字符串匹配以及随机回答来丰富机器人的形象。当获得信息后进入下一个状态。获得查询问题后给出结果。在询问股票时支持多种股票同时查询，运用向量机复数轮添加股票。

电话号码直接用正则表达式来捕捉。

对于三种信息，都在训练数据中分类，直接显示为不同的意图。用获取实体也行，但是 spacy 和向量机分类比较慢，不如直接意图识别。

所需要获取实体的只有股票。由于股票数量太多，也无法在训练集中修改直接获取。根据张老师的建议，先只找查询股票的意图，再用 spacy 分析句法，获得实体。同时有些股票名称含有特殊字符“-”，用正则表达式进行补充搜索。支持单个句子含有多个股票。

当获得股票实体后，还需要向 API 求证股票是否存在，不存在则捕捉错误信息，返回 unavailable 并进行特殊处理。

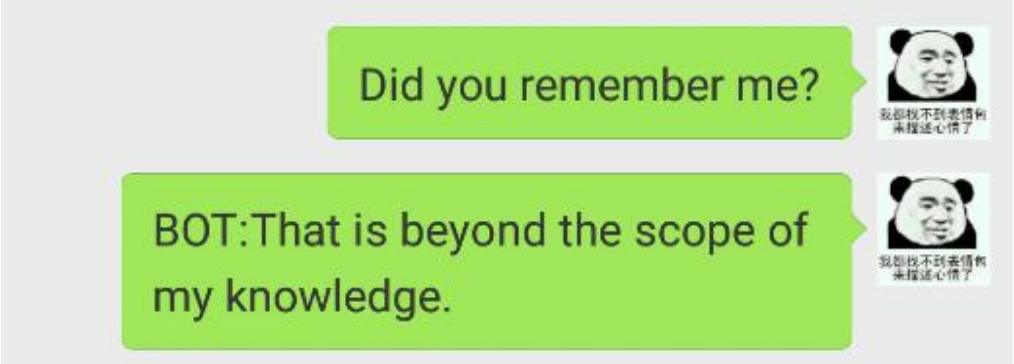
有些机器人的问句存在拒绝回答，同样运用意图识别来捕捉。

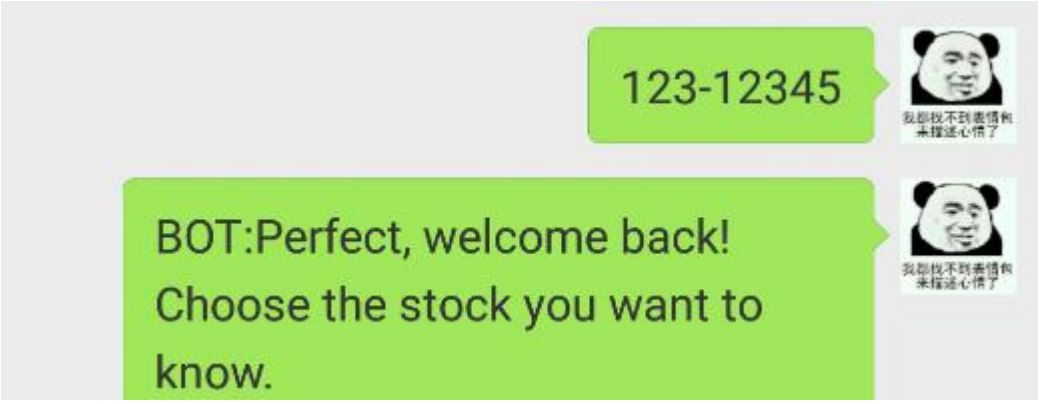
查询完毕后可以退出或者重新从选择股票开始，让状态机循环到之前的阶段。

三、成果展示

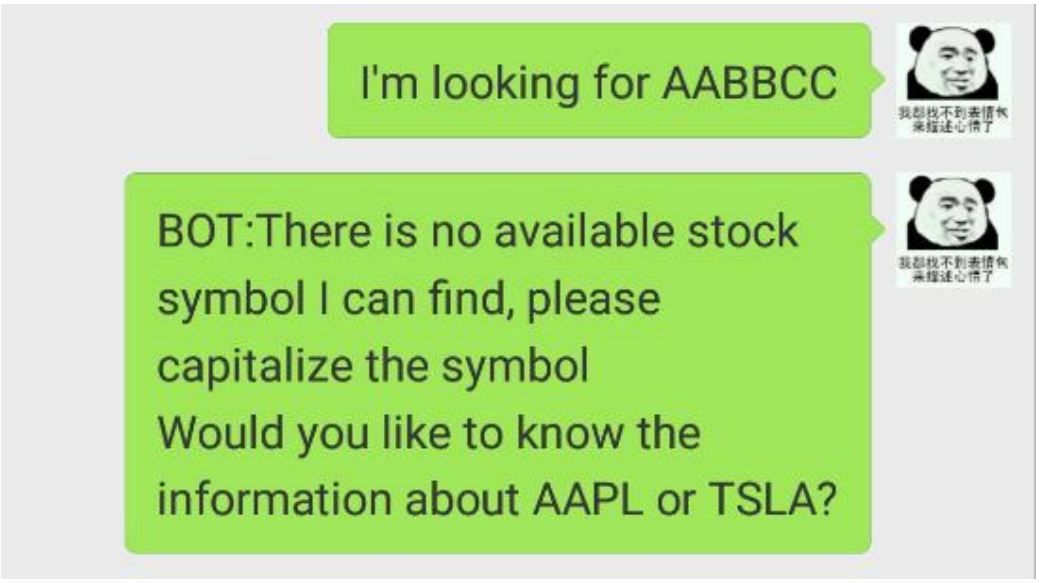
初始阶段，以及既定问题的回答，有随机多种方式。

Chitchat 以及之后用电话号码登录





查询股票，不存在的股票将会被提醒（AABBAA）



查询股票，存在多个股票时，不存在的股票（AABBAA）将不会被回答

Would you like to know the information about AAPL or TSLA?

AAPL

BOT:Perfect, trying to get the infromation
Would you like to know about the price or market capitalization?

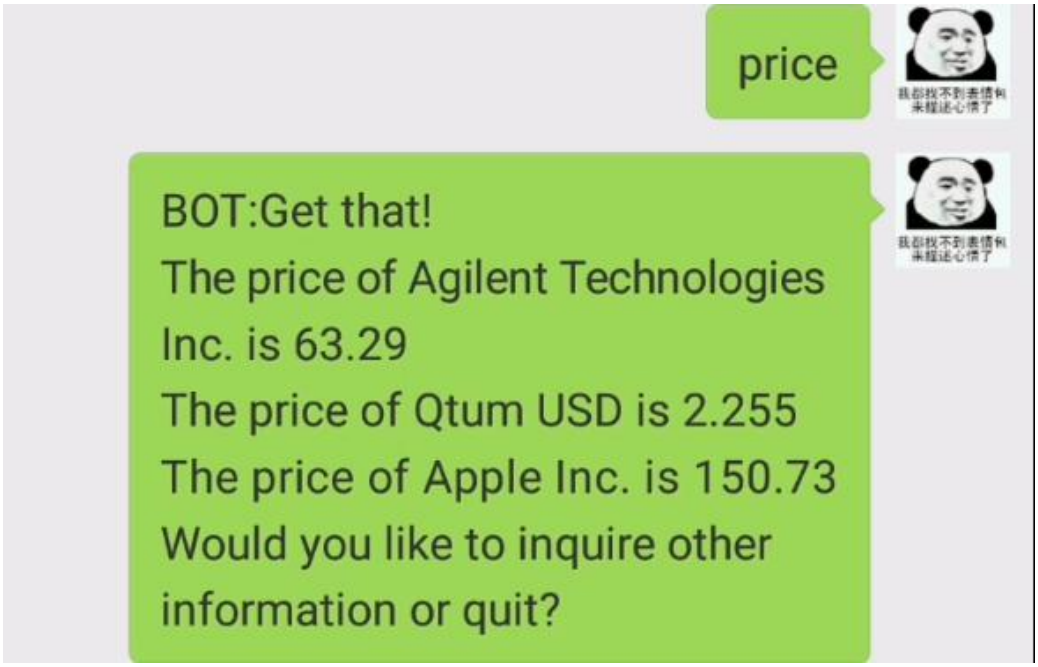
and AABBCC

BOT:Ok, add this to the stock list. What kind of information you want to know?

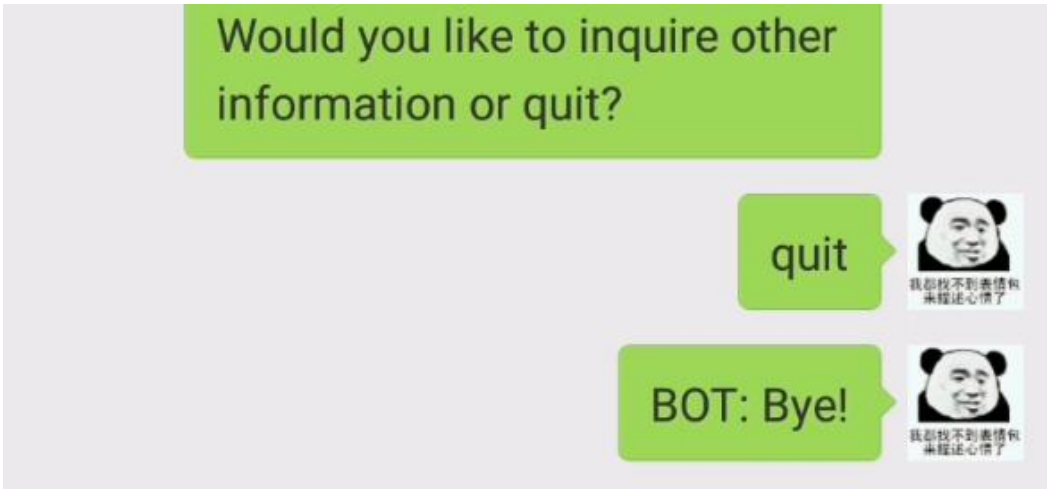
price

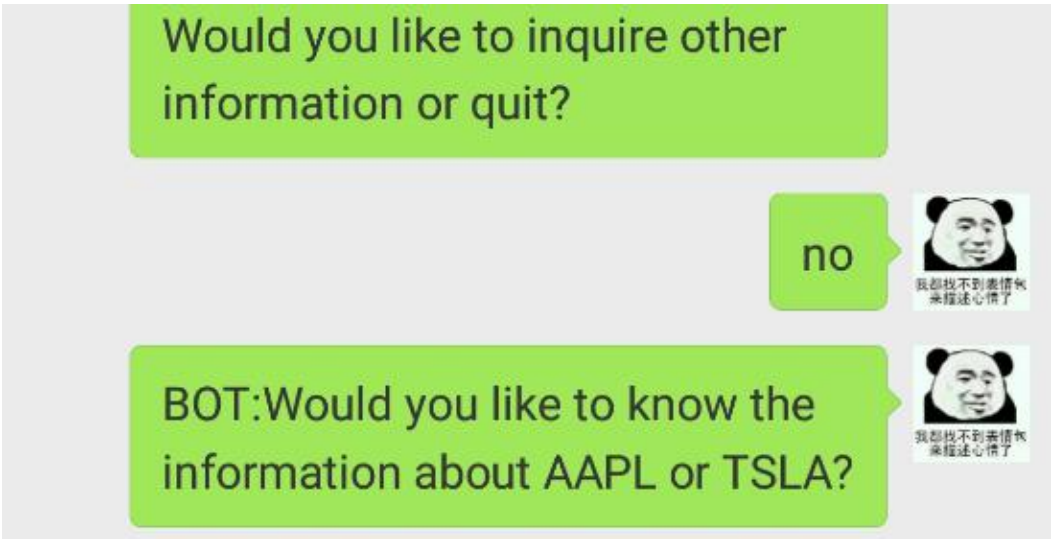
BOT:Get that!
The price of Apple Inc. is 150.73
Would you like to inquire other information or quit?

多个有效查询



退出或者循环到查询股票





状态机代码：

```
# Define the policy rules
policy_rules = {
    (INIT, "default"): (INIT, "You'll have to log in first, what's your phone number?", AUTHED),
    (INIT, "phone_number"): (AUTHED, "Perfect, welcome back! Choose the stock you want to know.", None),
    (AUTHED, "default"): (AUTHED, "Would you like to know the information about AAPL or TSLA?", CHOOSE_STOCK),
    (AUTHED, "negative"): (CHOOSE_OTHER_STOCK, "Ok, tell me the stock you want to choose", None),
    (AUTHED, "specify_stock"): (CHOOSE_STOCK, "Perfect, trying to get the infromation", CHOOSE_QUE),
    (CHOOSE_STOCK, "default"): (CHOOSE_STOCK,
                                "Would you like to know about the price or market capitalization?", CHOOSE_QUE),
    (CHOOSE_STOCK, "specify_stock"): (CHOOSE_STOCK,
                                      "Ok, add this to the stock list. What kind of information you want to know?",
                                      CHOOSE_QUE),
    (CHOOSE_STOCK, "negative"): (CHOOSE_OTHER_QUE, "Ok, then I guess you want to know the turnover, right?", None),
    (CHOOSE_STOCK, "specify_question"): (CHOOSE_QUE, "Get that!", None),
    (CHOOSE_OTHER_STOCK, "default"): (CHOOSE_OTHER_STOCK,
                                      "I didn't get that, tell me the stock you want to choose more clearly",
                                      CHOOSE_STOCK),
    (CHOOSE_OTHER_STOCK, "specify_stock"): (CHOOSE_STOCK, "Perfect, trying to get the infromation", CHOOSE_QUE),
    (CHOOSE_QUE, "specify_question"): (CHOOSE_QUIT, "", None),
    (CHOOSE_QUE, "specify_stock"): (CHOOSE_QUIT, "", None),
    (CHOOSE_QUE, "default"): (CHOOSE_QUIT, "", None),
    (CHOOSE_OTHER_QUE, "specify_question"): (CHOOSE_QUIT, "", None),
    (CHOOSE_OTHER_QUE, "default"): (CHOOSE_QUIT, "", None),
    (CHOOSE_QUIT, "default"): (CHOOSE_QUIT, "Would you like to inquire other information or quit?", CHOOSE_QUE),
    (CHOOSE_QUIT, "specify_question"): (CHOOSE_QUE, "OK", None),
    (CHOOSE_QUIT, "quit"): (INIT, "See you next time!", None),
    (CHOOSE_QUIT, "negative"): (AUTHED, "Would you like to know the information about AAPL or TSLA?", CHOOSE_STOCK)
}
```

语义识别以及获取实体：

```
I'm looking for AABBCC 1 None
{'intent': {'name': 'stock_search', 'confidence': 0.6475960969605173},
['AABBCC']
```

四、总结与感想

在这次项目中，我学习到了很多新的知识，并成功地将其运用到了聊天机器人中。从最开始的正则表达式到最后的状态机多轮多次查询，在老师的引导下，我们一步步了解了聊天机器人的构造。就像搭积木一般，每一次上课老师都教会我们一块积木，不知不觉间回头一看聊天机器人的基本框架已经构成了。这种高屋建瓴的方式让我感触颇深。

同时，老师也给了我们很大的自由空间，让我们去自主学习一些东西，这大大锻炼了我们的实际操作能力。不会用 API，就去查询它的文档。在遇到困难时，老师也会尽可能的帮助我。一开始我就不能很好地抽取股票实体，还是老师告诉我用 spacy，doc.ent 来捕捉能取到更好的效果。课上的小练习以及课后的复习指导都有效提高了我们的学习效率。

当然这次我的聊天机器人还有不足之处。比如上课学到的查询数据库就没能很好地结合到最终的项目中。训练数据也有不完善的地方，有时候会出现错误的语义识别。但总体而言，能完成一个可以有实际用处的聊天机器人还是让我有了巨大的成就感。

在参加这次项目之前，人工智能对于我来说仍然是个模糊的概念。虽然对其中部分算法有所听闻，却也是一知半解。老师由浅入深的讲解让我对人工智能有了一个整体的印象。刚好大四我也学习了几门人工智能的课程，老师之前的讲解对我帮助颇多。一门课的结课项目我也用了“识别面部情绪”这一老师提到过的研究作为课题，并成功获得了满分。这次的经历大大激发了我对人工智能的了解和兴趣，希望我能在这条道路上走的越来越远。