远程科研指导学习报告

* 智能聊天机器人

报告人：李宇泰

1、实验背景

2、主要知识点总结

2-1. 正则表达式

2-2. 状态机

2-3. RASA NLU

3、实验总结

## 1、实验背景

聊天机器人是一个用来模拟人类对话或聊天的程序，基础目标是至少暂时的让一个人认为他们正在与另一个人交谈。基本原理则是研发者将自己想要输出的回答放到数据库中，当用户输入问题时，程序根据算法从数据库中找到最贴切的答案，并将之回复给用户

与聊天机器人相关的人工智能领域是自然语言处理，通常“弱人工智能”(Weak AI) 领域使用专门的软件或编程语言以完成特定的功能。例如 A.L.I.C.E.使用一种叫做 AIML 的标记式语言，适用于谈话代理的功能，并且已被各类开发人员采用，他们的产品叫做爱丽丝机器人(Alicebots)。不过 A.L.I.C.E.仍是纯粹运用类型配对的技巧，缺乏思考能力， 跟1966年的 ELIZA 并无不同。“强人工智能”(strong AI) 则不同, 必须有智慧和逻辑推理的能力。

聊天机器人通常整合于对话系统，例如[虚拟助理](https://baike.baidu.com/item/%E8%99%9A%E6%8B%9F%E5%8A%A9%E7%90%86" \t "_blank)，使其能够在其主要的专家系统的范畴之外进行闲聊等等。

*通讯平台*

目前聊天机器人广泛运用于[即时通讯](https://baike.baidu.com/item/%E5%8D%B3%E6%97%B6%E9%80%9A%E8%AE%AF/111144" \t "_blank)平台，例如脸书 Messenger，WeChat，[LINE](https://baike.baidu.com/item/LINE)和 Kik，以娱乐、零售行销、以及客服为目的。此外，即时通讯平台提供易于整合的webhook，使得第三方开发商易于可通用于不同通讯平台之聊天机器人。这些软件机器人以客服的身份出现或是成为团体聊天的一员。有些即时通讯的机器人可以连接外部数据库，提供使用者新闻，气象，导航，电影放映时间，股价等资讯。达美乐、必胜客、迪士尼、Nerdify、雅玛多 Line、全食超市都已推出各自的聊天机器人，以便与终端消费者增进交流，推销公司的产品与服务，并且让消费者订货更加方便。2016年，观光业的一些旅行社和航空公司透过 Messenger 推出了聊天机器人的服务，墨西哥航空利用人工智能售票、回答问题，墨航和荷兰皇家航空并且提供航班资讯，处理乘客报到，发出行动登机证，推荐旅馆、餐厅、目的地行程。中国的旅行社在此之前已用 Wechat 提供这些服务。

*教育*

有些聊天机器人，例如 Nerdify 开发的 Nerdy Bot，针对大中小学生面对的问题，让学习更简单又有效率。该软件利用脸书 Messenger 即时回答学生作业相关的问题以便加速学习。加大尔湾分校图书馆的聊天机器人 ANTswers，2014年开始试用，被认为非常成功。 [2]

从旁观者的角度可能惠人为一个机器人能够回答的问题越多，则越为智能，可是事实上只要有着足够数量的数据，无论什么算法或者结构都能在回答问题这个方面达到相似的正确率，所以能够回答的问题数量仅仅代表这个程序背后的数据库大小。而智能并不是比较数据库的大小，而是更多方面的统一体现：

1. 学习能力

这是最根本去又是最难以提升的一条标准，一个可以根据用户的输入自动成长的程序往往是比一个有着庞大数据库却不能成长的程序好的。因为它的本意是帮助人类，减轻人类在各个方面的劳动量，可如果一个程序或是机器人需要手动录入所有的问题，这本身就增加了人类的劳动量。这也是为何聊天机器人虽然热门，却应用较少的一个根本原因

2. 数据筛选能力

在一个程序拥有学习能力之后，记忆不评判机器人的智能程度的一个标准是机器人是否有着选择性的学习，如果机器人智能全盘接受而不嗯呢该对信息进行筛选的话，这个学习能力时并不完整的。

3. 自升级能力

在程序能狗进行自我学习并且能够对数据进行一定程度上的筛选之后，自升级能力是下一个壁垒，这个部分的主旨在于如何让机器人尝试掉正算法规则和数据组织结构一时的在误率较高的知识组织形式或领域降低学习的错误率。

## 2、主要知识点回顾

### 1. 正则表达式

又称规则表达式，计算机科学的一个概念，常常用力啊检索、替换符合某种模式（规则）的文本。正则表达式是对字符串（包括普通字符（例如，a 到 z 之间的字母）和特殊字符（称为“元字符”））操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种文本模式，该模式描述在搜索文本时要匹配的一个或多个字符串。

对于一个给定的正则表达式，我们可以达到一下目的：

1. 给定的字符串是否符合正则表达式的过滤逻辑

2. 通过正则表达式来从字符串中获取我们想要的部分

正则表达式的“鼻祖”或许可一直追溯到科学家对人类神经系统工作原理的早期研究。美国[新泽西州](https://baike.baidu.com/item/%E6%96%B0%E6%B3%BD%E8%A5%BF%E5%B7%9E" \t "_blank)的Warren McCulloch和出生在美国底特律的Walter Pitts这两位神经生理方面的科学家，研究出了一种用数学方式来描述神经网络的新方法，他们创造性地将神经系统中的神经元描述成了小而简单的自动控制元，从而作出了一项伟大的工作革新。

在1951 年,一位名叫Stephen Kleene的数学科学家，他在Warren McCulloch和Walter Pitts早期工作的基础之上，发表了一篇题目是《神经网事件的表示法》的论文，利用称之为正则集合的数学符号来描述此模型，引入了正则表达式的概念。正则表达式被作为用来描述其称之为“正则集的代数”的一种表达式，因而采用了“正则表达式”这个术语。之后一段时间，人们发现可以将这一工作成果应用于其他方面。Ken Thompson就把这一成果应用于计算[搜索算法](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E7%AE%97%E6%B3%95" \t "_blank)的一些早期研究，Ken Thompson是 Unix的主要发明人，也就是大名鼎鼎的Unix之父。Unix之父将此[符号系统](https://baike.baidu.com/item/%E7%AC%A6%E5%8F%B7%E7%B3%BB%E7%BB%9F" \t "_blank)引入编辑器QED，然后是Unix上的编辑器ed，并最终引入grep。Jeffrey Friedl 在其著作《Mastering Regular Expressions (2nd edition)》（中文版译作：精通正则表达式，已出到第三版）中对此作了进一步阐述讲解，如果你希望更多了解正则表达式理论和历史，推荐你看看这本书。

自此以后，正则表达式被广泛地应用到各种UNIX或类似于UNIX的工具中，如大家熟知的[Perl](https://baike.baidu.com/item/Perl" \t "_blank)。Perl的正则表达式源自于Henry Spencer编写的regex，之后已演化成了pcre（Perl兼容正则表达式Perl Compatible Regular Expressions），pcre是一个由Philip Hazel开发的、为很多现代工具所使用的库。正则表达式的第一个实用应用程序即为Unix中的 qed 编辑器。

然后，正则表达式在各种计算机语言或各种应用领域得到了广大的应用和发展

在最近的六十年中，正则表达式逐渐从模糊而深奥的数学概念，发展成为在计算机各类工具和软件包应用中的主要功能。不仅仅众多UNIX工具支持正则表达式，近二十年来，在WINDOWS的阵营下，正则表达式的思想和应用在大部分 Windows 开发者工具包中得到支持和嵌入应用！从正则式在Microsoft Visual Basic 6 或 Microsoft VBScript到.NET Framework中的探索和发展，WINDOWS系列产品对正则表达式的支持发展到无与伦比的高度，几乎所有 Microsoft 开发者和所有.NET语言都可以使用正则表达式。如果你是一位接触计算机语言的工作者，那么你会在主流操作系统（\*nix[Linux, Unix等]、Windows、HP、BeOS等）、主流的开发语言（delphi、[Scala](https://baike.baidu.com/item/Scala" \t "_blank)、[PHP](https://baike.baidu.com/item/PHP" \t "_blank)、[C#](https://baike.baidu.com/item/C%23" \t "_blank)、Java、[C++](https://baike.baidu.com/item/C%2B%2B" \t "_blank)、Objective-c、Swift、VB、[Javascript](https://baike.baidu.com/item/Javascript" \t "_blank)、[Ruby](https://baike.baidu.com/item/Ruby" \t "_blank)以及[Python](https://baike.baidu.com/item/Python" \t "_blank)等）、数以亿万计的各种应用软件中，都可以看到正则表达式优美的舞姿。

在本次实验中，利用了正则表达式来确定用户的输入是否符合标准，以此为基础创建了一个简单的login系统

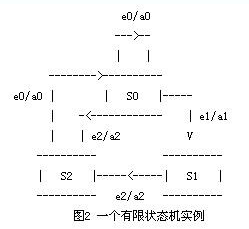
并且在特定的时候从用户的输入中提取出问题所需要的主体。

### 

### 2. 状态机

### 这次我们实验中所用到的状态机应该为有限状态机，思想广泛应用于硬件控制电路设计，也是软件上常用的一种处理方法。它把复杂的控制逻辑分解成有限个稳定状态，在每个状态上判断事件，变连续处理为离散数字处理，符合计算机的工作特点。同时，因为有限状态机具有有限个状态，所以可以在实际的工程上实现。但这并不意味着其只能进行有限次的处理，相反，有限状态机是闭环系统，有限无穷，可以用有限的状态，处理无穷的事务。../状态机1.jpg

有限状态机的工作原理如图1所示，发生事件(event)后，根据当前状态(current state) ，决定执行的动作(action)，并设置下一个状态号(next state)。



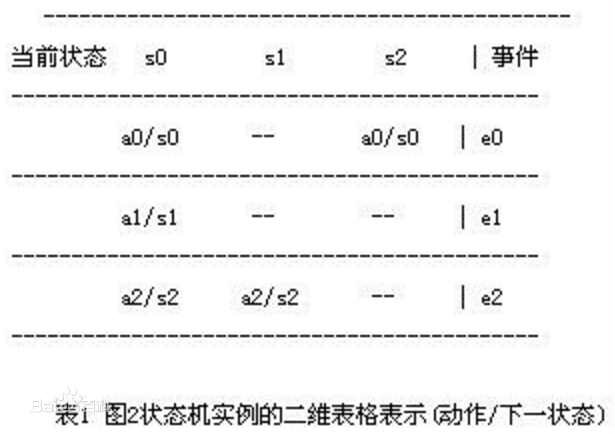
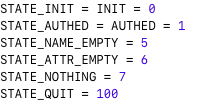


表1图2为一个状态机实例的状态转移图，它的含义是：在s0状态，如果发生e0事件，那么就执行a0动作，并保持状态不变；如果发生e1事件，那么就执行a1动作，并将状态转移到s1态；如果发生e2事件，那么就执行a2动作，并将状态转移到s2态；在s1状态，如果发生e2事件，那么就执行a2动作，并将状态转移到s2态；在s2状态，如果发生e0事件，那么就执行a0动作，并将状态转移到s0态

此次试验中，程序主体依赖于状态机而运行，程序中的状态机分为6个状态，分别为：

1.程序运行开始并输出提示信息

../../state%201.png

2.成功进行登录操作，输出提示并开始第一轮询问

../state2.png

3.缺少股票姓名时的再次询问../state%203.png

4.缺少股票查询内容是的再次询问

../state%204.png

5.输入信息完全不匹配当前所需要内容

../state%205.png

6.用户所需信息获取完毕，退出程序

../state%206.png

### 3.Rasa NLU

在rasa nlu中，数据分为意图、实体和置信度

意图即是用户的诉求，例如投诉，查询和要求退款等请求；而实体则是永福诉求细节的属性；置信度是一个距离指标，能够体现出nlu分析出的结果与意图列表中的诉求相差多少

此外，在数据库内，对于数据也有着一套分类方法

文本，意图以及实体类别



## 3.实验总结

本次科研项目虽然只有一个月，但是我从中接触到了太多的新知识、新概念，每一次交流都能学到新的东西，每一小部分对我来说都是一次小小挑战，在这段时间里我也学会了很多python的知识，对于聊天机器人也有了更深的理解。也是因为我的python基础并不是很好，所以在项目完成的过程中遇到非常多的问题，而因为我的基础并不是很牢固，所以我往往不知道问题出在哪里，也不能一而再再而三的麻烦老师，只能自己一边学习python一边查错，虽然这个过程稍长了些，但是我所收获的，却是非常所令人振奋的；另一方面在这次实验也锻炼了我独立作业的能力，以往在学校里经常会和同学一起讨论，任务完成了，却得不到很好的锻炼，而这次实验在一定程度上锻炼了了我的个人能力。

就本次实验而言，谈不上很成功，但也算是完成了任务，非常感谢老师给我这次机会，也感谢老师的耐心。