第3课 机器学习构建chatbot

寒小阳

2017.04.15

主要内容

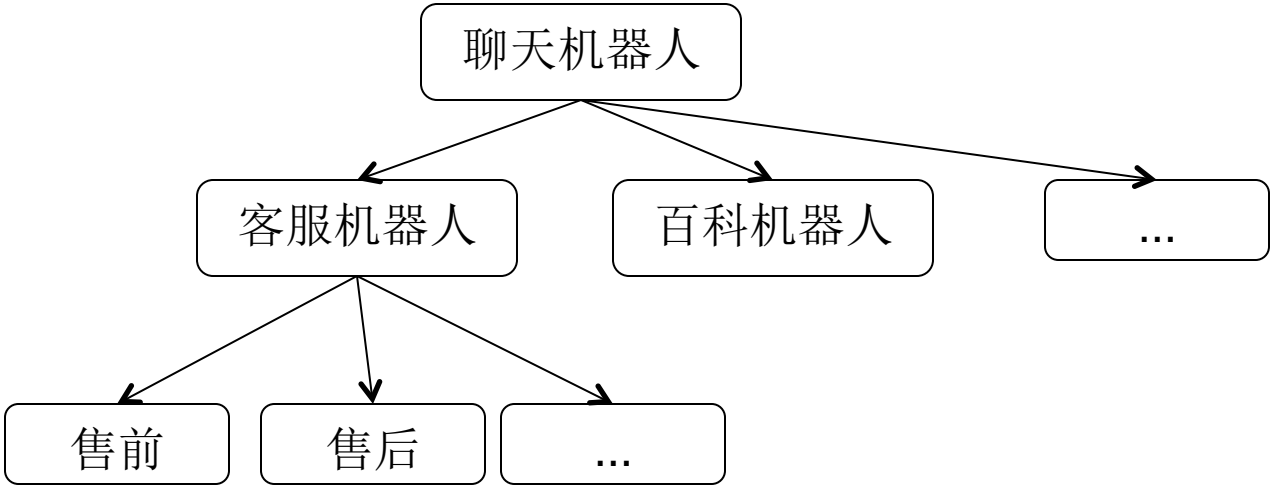
- 关于聊天机器人的思考
  1. 工程考量
  2. 机器学习角度考虑
- 预备知识
  1. 检索与匹配
  2. 分类与朴素贝叶斯
- chatterbot
  1. 架构与使用方法
  2. 源码分析

传统聊天机器人

- NLP基础知识
  - 基本分词
  - 关键词抽取(tf-idf等)
  - 正则表达式模式匹配
  - …
- Machine Learning相关知识
  - 文本表示与匹配
  - 分类(文本场景分析)
  - 数据驱动(特征工程)
  - …

聊天机器人的一些思考

- 工程考量
  - 架构设计清晰、模块化
  - 功能分析，解耦，部件可插拔与扩展



聊天机器人的一些思考

- 算法与机器学习角度考量
  - 算法简单，数据(特征)驱动
  - 场景化与垂直领域

预备知识：检索与匹配

- 基于检索与匹配
  - 知识库（存储了问题与回复内容）
  - 检索：搜寻相关问题
  - 匹配：对结果进行排序

预备知识：编辑距离

编辑距离

- 编辑距离/Levenshtein距离，是指两个字符串之间，由一个转成另一个所需要的最少编辑操作次数。
- 允许的编辑操作包括：
  - 将一个字符替换成另一个字符
  - 插入一个字符
  - 删除一个字符

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

- 初始化：

$$D(i,0) = i$$
$$D(0,j) = j$$

- 递归方程:

$$D(i,j) = \begin{cases} D(i-1,j) + 1 \\ \min\{D(i,j-1)\} + 1 \\ D(i-1,j-1) + \begin{cases} 2; \text{ if } X(i) \neq Y(j) \\ 0; \text{ if } X(i) = Y(j) \end{cases} \end{cases}$$

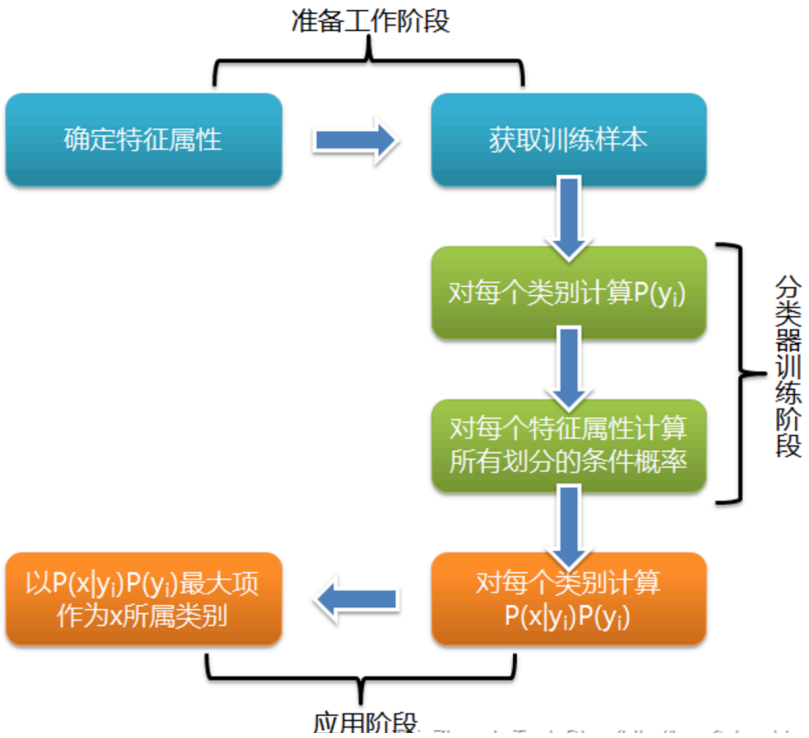
预备知识：python编辑距离

- Python在string 类型中，默认的 utf-8 编码下，一个中文字符是用三个字节来表示的。用unicode。

```
# -*- coding:utf-8 -*-
import Levenshtein
texta = u'七月在线'
textb = u'七月·在线'
print Levenshtein.distance(texta,textb)
```

预备知识：场景分类与NB

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$



Chatterbot聊天机器人

ChatterBot是一个基于机器学习的聊天机器人引擎，构建在python上，主要特点是可以自可以从已有的对话中进行学(jiyi)习(pipei)。

Chatterbot聊天机器人

- 每个部分都设计了不同的“适配器”(Adapter)
  - 机器人应答逻辑 => Logic Adapters
    - Closest Match Adapter
      - 字符串模糊匹配(编辑距离)
    - Closest Meaning Adapter
      - 借助nltk的WordNet，近义词评估
    - Time Logic Adapter
      - 处理涉及时间的提问
    - Mathematical Evaluation Adapter
      - 涉及数学运算

Chatterbot聊天机器人

- 聊天机器人应答逻辑请参考课程代码

Chatterbot聊天机器人

- 每个部分都设计了不同的“适配器”(Adapter)
  - 存储器后端 => Storage Adapters
    - Read Only Mode
      - 只读模式，当有输入数据到chatterbot的时候，数据库并不会发生改变
    - Json Database Adapter
      - 用以存储对话数据的接口，对话数据以Json格式进行存储。
    - Mongo Database Adapter
      - 以MongoDB database方式来存储对话数据

Chatterbot聊天机器人

- 聊天机器人存储适配器请参考课程代码

Chatterbot聊天机器人

- 每个部分都设计了不同的“适配器”(Adapter)
  - 输入形式 => Input Adapters
    - Variable input type adapter
      - 允许chatter bot接收不同类型的输入的，如strings,dictionaries和Statements
    - Terminal adapter
      - 使得ChatterBot可以通过终端进行对话
    - HipChat Adapter
      - 使得ChatterBot 可以从HipChat聊天室获取输入语句，通过HipChat 和 ChatterBot 进行对话
    - Speech recognition
      - 语音识别输入，详见chatterbot-voice

Chatterbot聊天机器人

- 聊天机器人输入适配器请参考课程代码

Chatterbot聊天机器人

- 每个部分都设计了不同的“适配器”(Adapter)
  - 输出形式 => Output Adapters
    - Output format adapter
      - 支持text，json和object格式的输出
    - Terminal adapter
    - HipChat Adapter
    - Mailgun adapter
      - 允许chat bot基于Mailgun API进行邮件的发送
    - Speech synthesis
      - TTS(Text to speech)部分，详见chatterbot-voice

Chatterbot源码解析

请参考课上讲解部分

感谢大家么么哒！

恳请大家批评指正！