

*建议本文档使用markdown编辑器打开

介绍

本项目的核心部分是基于tf-idf检索的召回模型，构建召回+排序的客服聊天机器人。系统支持FAQ问答模式的客服机器人，采取的数据集是小鸡孵化器相关垂直领域的FAQ问答数据集。

目前该系统的优点在于：

- 一、召回+排序 2个模块互不干扰，便于自定义修改以及维护；
- 二、系统采取了排序规则优化，提升了检索速度。
- 三、加入了简单的倒排索引，优化了检索流程。

环境配置

Python版本为3.6

需要创建Python=3.6的虚拟环境

详细配置见requirements.txt或者qa.yaml

这里提供两种配置项目运行环境的方法

1.使用pip安装

```
conda create -n qa python=3.6 #创建名为qa, Python版本为3.6的conda虚拟环境

conda activate qa # 激活虚拟环境

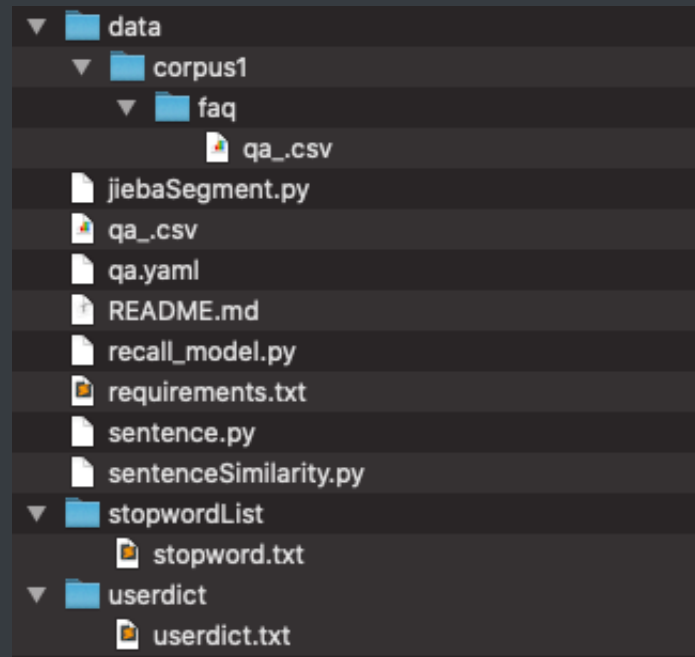
pip install -r requirements.txt # 安装环境依赖库
```

2.使用conda安装

```
conda env create -f env.yaml
```

个人建议：更换清华源后，使用第一种方法安装

文件结构说明



qa.csv文件为数据存放文件

即孵化器问答对

工程师如需添加问答对，直接在文件末尾添加即可

jiebaSegment.py、**sentence.py**、**sentenceSimilarity.py**均为辅助文件

里面包含了主文件运行需要调用的、且开源库没有的自定义函数

stopword.txt文件为项目停用词一般不需要工程师修改

userdict.txt文件为用户定义词典

为了预防数据集中出现生僻词，jieba分词不可识别，这时需要工程师自行添加

里面已经包含了示例

项目运行

进入项目文件夹顶层

激活conda环境

```
conda activate qa
```

运行主函数文件

```
python recall_model.py
```

部分代码说明

```
24 def read_corpus1():
25     qList = []
26     # 问题的关键词列表
27     qList_kw = []
28     aList = []
29     data = pd.read_csv('./data/corpus1/faq/qa_.csv', header=None)#读取CSV文件
30     data_ls = np.array(data).tolist()
31     for t in data_ls:
32         qList.append(t[0])
33         qList_kw.append(seg.cut(t[0]))
34         aList.append(t[1])
35     return qList_kw, qList, aList
```

recall_model.py文件第29行：读取问答对

```
113 if __name__ == '__main__':
114     # 设置外部词
115     seg = Seg()
116     seg.load_userdict('./userdict/userdict.txt')
```

recall_model.py文件第116行：读取外部词

```
128 while True:
129     question = input("请输入问题(q退出): ")
130     if question == 'q':
131         break
132     time1 = time.time()
133     question_k = ss.similarity_k(question, 5)
134     print("您好, 我给您找到的答案是: {}".format(answerList[question_k[0][0]]))
135     for idx, score in zip(*question_k):
136         print("same questions: {}, score: {}".format(questionList[idx], score))
137     time2 = time.time()
```

recall_model.py文件第129行：输入问题

recall_model.py文件第129行：输出匹配的问题和回答

```
123     ss.set_sentences(questionList)
124     ss.TfidfModel()           # tfidf模型
125     # ss.LsiModel()           # lsi模型
126     # ss.LdaModel()           # lda模型
```

recall_model.py文件第124行：选择调用核心函数

三个召回模型已经在函数文件写好，在主文件直接调用即可

tfidf模型，精度最高，但是语义映射到了高维空间，速度相对其他两个较慢

建议问答对5位数以内使用该模型，5位数以上使用其他两个

运行Demo

```

1. mazhanyu@mzy-MacBook-Pro:~/Desktop/test (zsh)
# mazhanyu @ mzy-MacBook-Pro in ~/Desktop/test [23:47:07]
$ python recall_model.py
/Users/mazhanyu/anaconda3/envs/qa/lib/python3.6/site-packages/gensim/similarities/__init__.py:15: UserWarning: The
gensim.similarities.levenshtein submodule is disabled, because the optional Levenshtein package <https://pypi.org/p
roject/python-Levenshtein/> is unavailable. Install Levenshtein (e.g. `pip install python-Levenshtein`) to suppress
this warning.
  warnings.warn(msg)
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ky/k7zxc4cj29x_q5lt4lqgxfkm0000gn/T/jieba.cache
Loading model cost 0.720 seconds.
Prefix dict has been built successfully.
<sentenceSimilarity.SentenceSimilarity object at 0x7f9203e84898>
请输入问题(q退出): 胚胎如何吸收营养
您好, 我给您找到的答案是: 胚胎发育当中需要的营养物质, 大部分是由蛋黄提供; 而水份是由蛋清提供; 钙质及微量元素由蛋壳
提供; 空气交换由气室实现。而这些物质元素是如何进入胚胎体内的呢? 其实主要靠的就是血管来进行运输, 血液流通的动力源自
心脏的跳动。所以, 心脏是最先发育的器官, 当我们照蛋的时候就会发现, 心脏跳动后, 血管就会不断地生长, 不断的获取营养物
质。
same questions: 胚胎如何吸收营养?, score: 1.0
same questions: 胚胎怎样呼吸?, score: 0.23434005677700043
same questions: 孵化期间, 胚胎是如何呼吸?, score: 0.18194854259490967
same questions: 胚胎如何补充钙质?, score: 0.152542382478714
same questions: 孵化前的准备?, score: 0.0
Time cost: 0.001177072525024414 s
请输入问题(q退出): 他怎么吸收营养
您好, 我给您找到的答案是: 胚胎发育当中需要的营养物质, 大部分是由蛋黄提供; 而水份是由蛋清提供; 钙质及微量元素由蛋壳
提供; 空气交换由气室实现。而这些物质元素是如何进入胚胎体内的呢? 其实主要靠的就是血管来进行运输, 血液流通的动力源自
心脏的跳动。所以, 心脏是最先发育的器官, 当我们照蛋的时候就会发现, 心脏跳动后, 血管就会不断地生长, 不断的获取营养物
质。
same questions: 胚胎如何吸收营养?, score: 0.9205745458602905
same questions: 孵化前的准备?, score: 0.0
same questions: 要怎么孵化?, score: 0.0
same questions: 要怎么玩呢?, score: 0.0
same questions: 蛋壳有什么作用?, score: 0.0
Time cost: 0.0007648468017578125 s
请输入问题(q退出): 胚胎要怎么吸收营养
您好, 我给您找到的答案是: 胚胎发育当中需要的营养物质, 大部分是由蛋黄提供; 而水份是由蛋清提供; 钙质及微量元素由蛋壳
提供; 空气交换由气室实现。而这些物质元素是如何进入胚胎体内的呢? 其实主要靠的就是血管来进行运输, 血液流通的动力源自
心脏的跳动。所以, 心脏是最先发育的器官, 当我们照蛋的时候就会发现, 心脏跳动后, 血管就会不断地生长, 不断的获取营养物
质。
same questions: 胚胎如何吸收营养?, score: 1.0
same questions: 胚胎怎样呼吸?, score: 0.23434005677700043
same questions: 孵化期间, 胚胎是如何呼吸?, score: 0.18194854259490967
same questions: 胚胎如何补充钙质?, score: 0.152542382478714
same questions: 孵化前的准备?, score: 0.0
Time cost: 0.0006029605865478516 s
请输入问题(q退出): 胚胎如何汲取营养
您好, 我给您找到的答案是: 胚胎发育当中需要的营养物质, 大部分是由蛋黄提供; 而水份是由蛋清提供; 钙质及微量元素由蛋壳
提供; 空气交换由气室实现。而这些物质元素是如何进入胚胎体内的呢? 其实主要靠的就是血管来进行运输, 血液流通的动力源自
心脏的跳动。所以, 心脏是最先发育的器官, 当我们照蛋的时候就会发现, 心脏跳动后, 血管就会不断地生长, 不断的获取营养物
质。
same questions: 胚胎如何吸收营养?, score: 0.7591252326965332
same questions: 胚胎怎样呼吸?, score: 0.3086974620819092
same questions: 孵化期间, 胚胎是如何呼吸?, score: 0.23968182504177094
same questions: 胚胎如何补充钙质?, score: 0.2009449154138565
same questions: 孵化前的准备?, score: 0.0
Time cost: 0.00080108642578125 s
请输入问题(q退出): q
(qa)
# mazhanyu @ mzy-MacBook-Pro in ~/Desktop/test [23:48:50]
$ 

```

可以看到最后的语义理解还是非常喜人的