


NLPartners - 对话机器人 设计



----2020.06.27



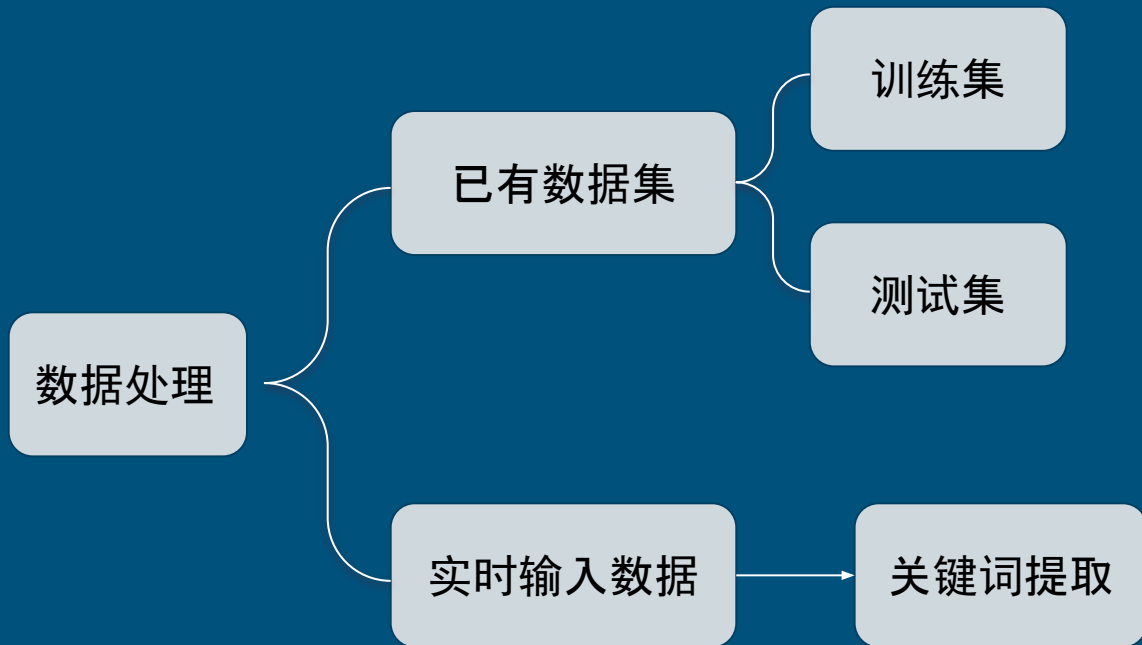
目录

1. 项目介绍
2. 数据处理
3. 模型设计
4. 项目效果展示

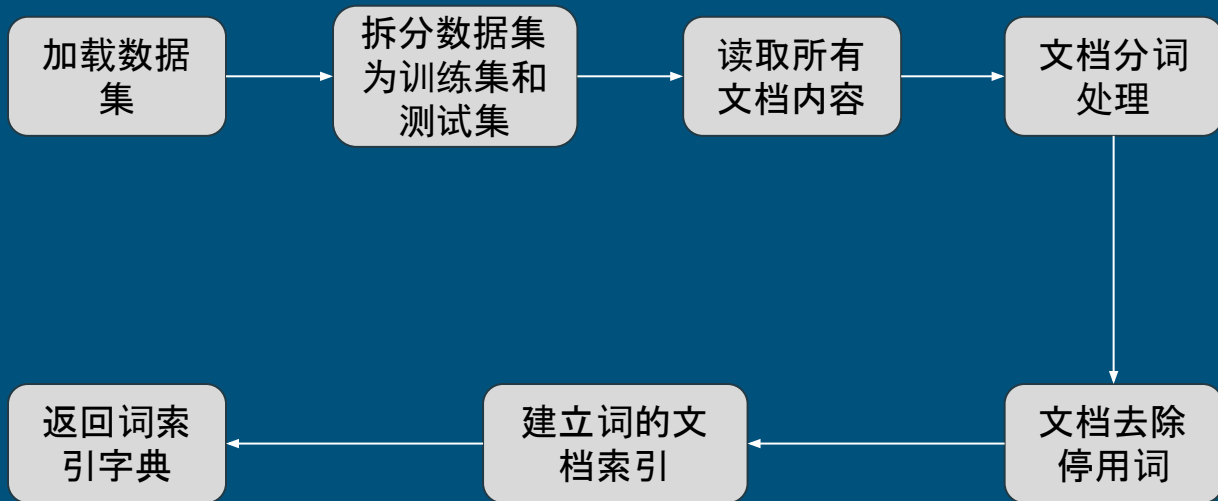
项目介绍

聊天机器人，是目前NLP重要应用场景之一，其又可细分为任务导向型，闲聊型和问答型三类，三类相互之间并没有明确的边界。本项目中设计的智能对话机器人，属于问答类型，主要用于提供银行业务相关的咨询服务。用户可在直接在页面中输入想要咨询的业务相关的问题，机器人将分析问题，然后提供相应的答案，并显示在对话框中，用户可及时地获取到想要的答案。对常见的业务，采用智能机器人的方式，可大大减少人工的重复劳动，提高了银行业务的处理效率。本项目的目的也即是对这类业务的简单实现。

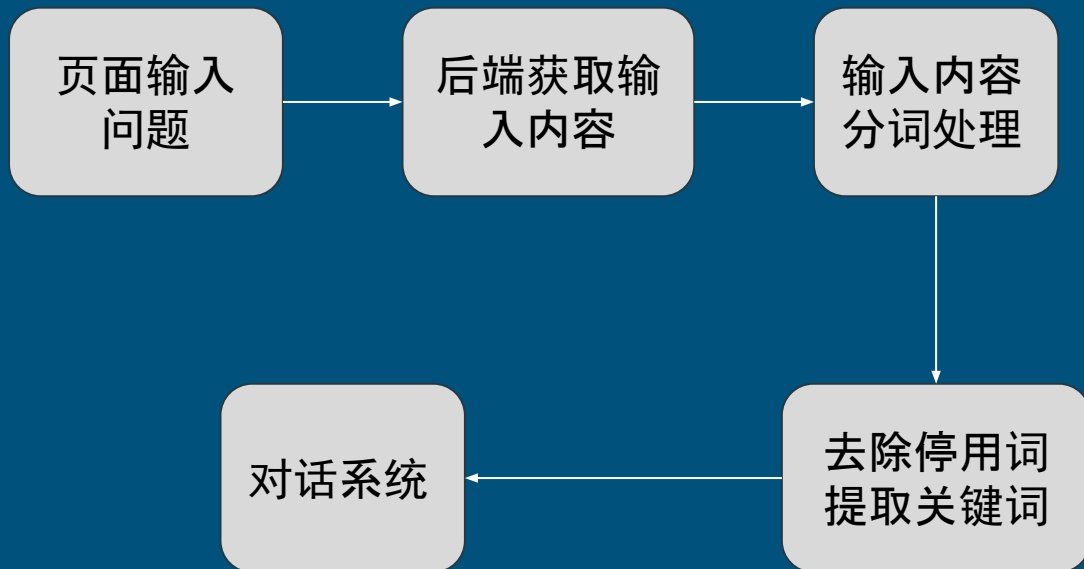
数据处理



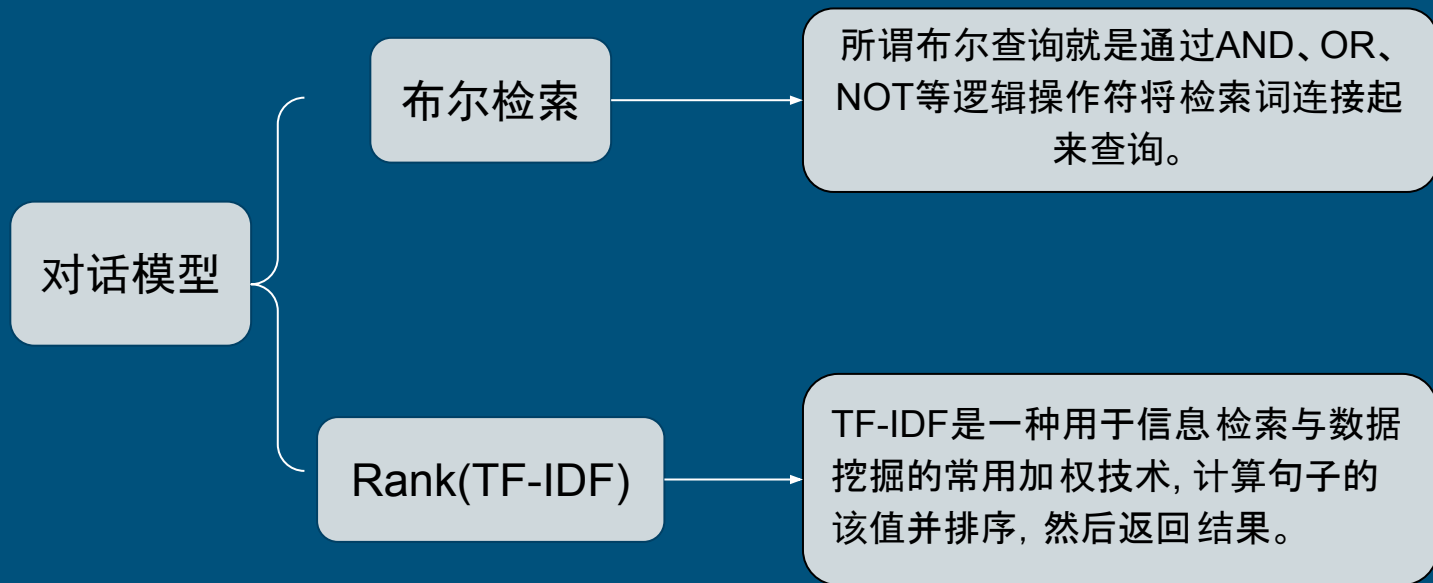
已有数据集处理



实时输入数据处理

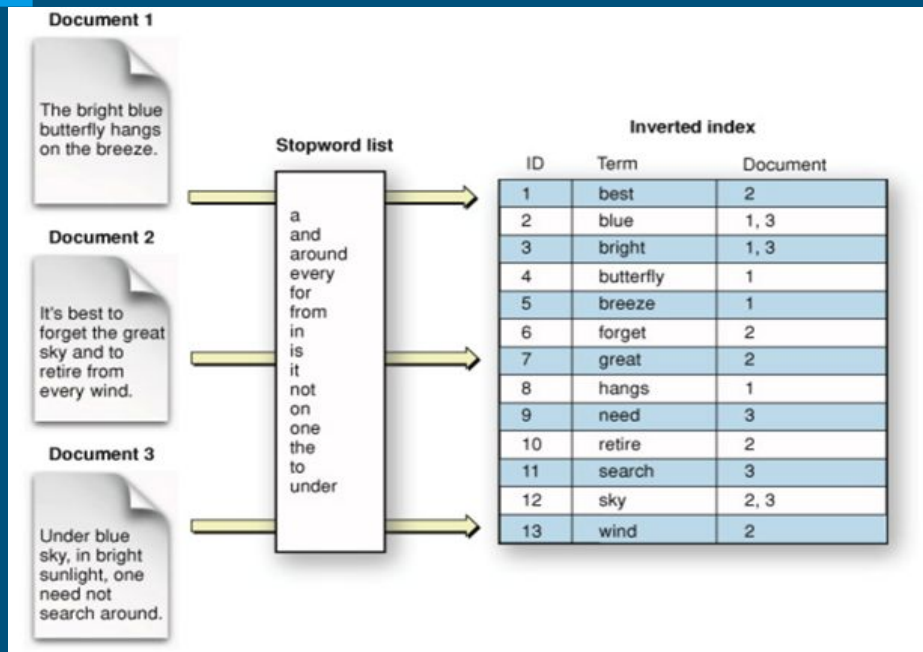


模型设计



本项目中采用布尔搜索和根据TF-IDF排序的方式, 来获取输入问题的结果。

布尔搜索



在布尔检索前需先构建好文档的倒排索引列表。检索某句话时，提取出关键词。如若提取出的是左图中“bright”和“blue”两个词，则有条件：“bright” and “blue”，即检索可知两个词同时出现在文档1中，则返回文档1的内容。当所得的结果是多个文档时，则还需最结果进行排序，返回最符合的结果。

收集需要建立索引的文档

分词并去停用词

倒排索引列表

关键词为“And”条件检索

TF-IDF(词频-逆文件频率)

词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。

逆向文件频率 (inverse document frequency, IDF) : 如果包含词条t的文档越少, IDF越大, 则说明词条具有很好的类别区分能力。

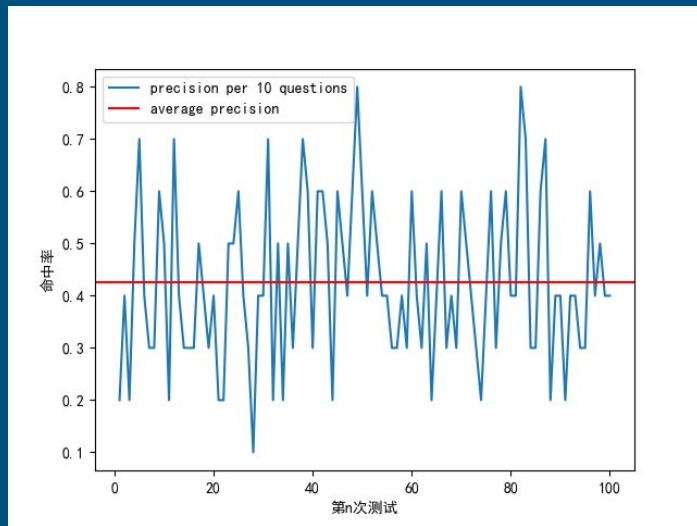
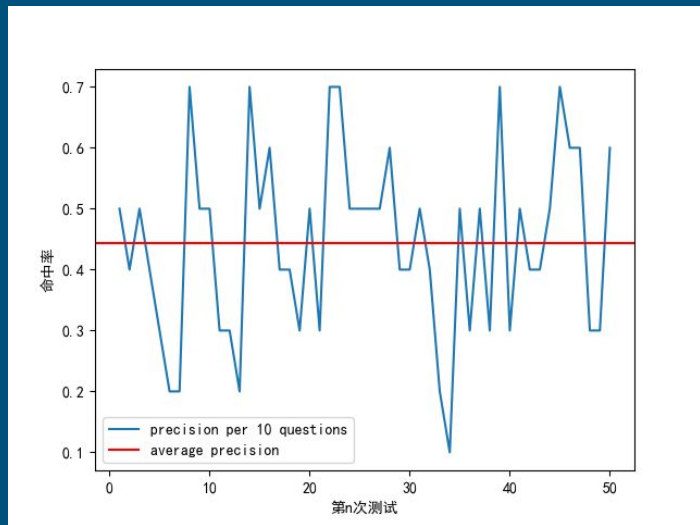
TF-IDF倾向于过滤掉常见的词语, 保留重要的词语。

$$TF_w = \frac{\text{在某一类中词条 } w \text{ 出现的次数}}{\text{该类中所有的词条数目}}$$

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条 } w \text{ 的文档数} + 1}\right),$$

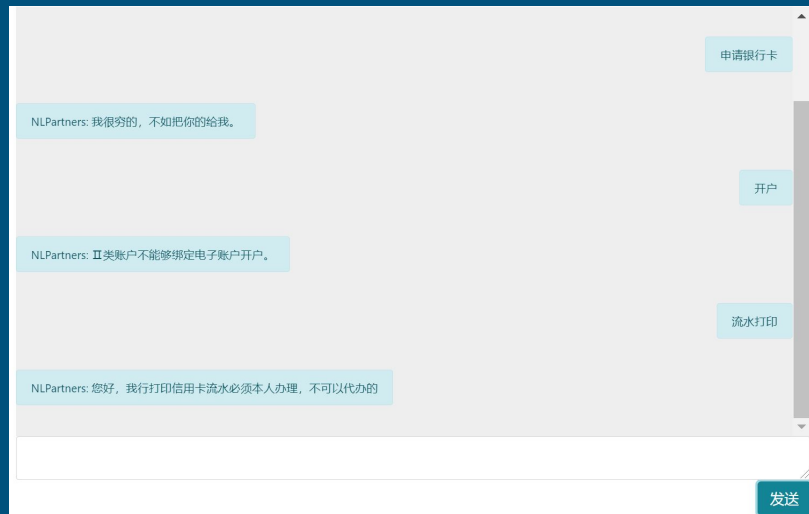
$$TF_IDF = TF * IDF$$

模型评估



利用测试集对模型的命中率进行评估，模型平均的命中率为45%左右。影响命中率的主要是

项目效果展示



项目优缺点说明

优点：

1. 采用布尔检索和TF-IDF结合的方式，系统设计简单，易于实现和操作；
2. 实现了基本的对话任务，输入内容后，可实时返回结果。
3. 页面简洁易操作；

缺点：

1. 语义分析方面还有待改善；
2. 语料库还有待完善，部分搜索没有结果。

谢谢