**基于深度学习的自然语言处理**

**第8次实验报告**

**一、实验名称**

实现基于知识图谱问答的整个流程

**二、实验日期**

2021.1.8 – 2021.1.22

**三、实验目的**

（1）熟悉知识图谱

（2）熟悉基于知识图谱问答

**四、实验数据**

提供的开源的实验数据

**五、实验步骤及结果分析**

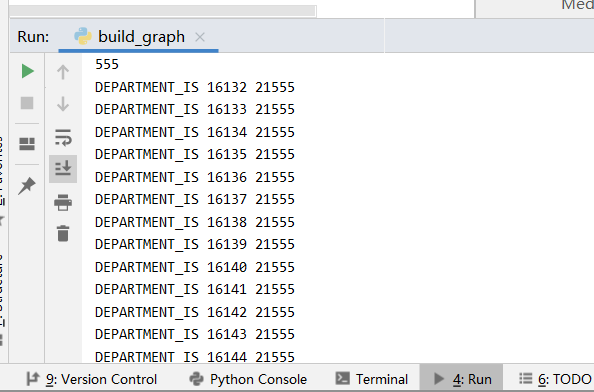
先下载词向量 等到data

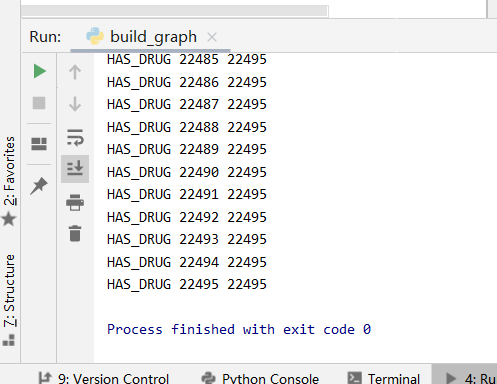
5.1搭建知识图谱：python build\_grapy.py。

1.neo4j.bat stop停止neo4j服务

2.neo4k.bat console 启动neo4j

3.搭建知识图谱python build\_grapy.py





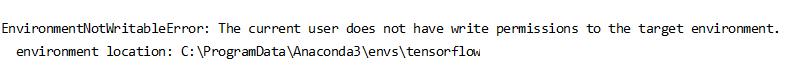
4.启动问答测试：python kbqa\_test.py

5.pip install pyahocorasick报错



改为使用

conda install -c https://conda.anaconda.org/conda-forge pyahocorasick



运行，依然报错



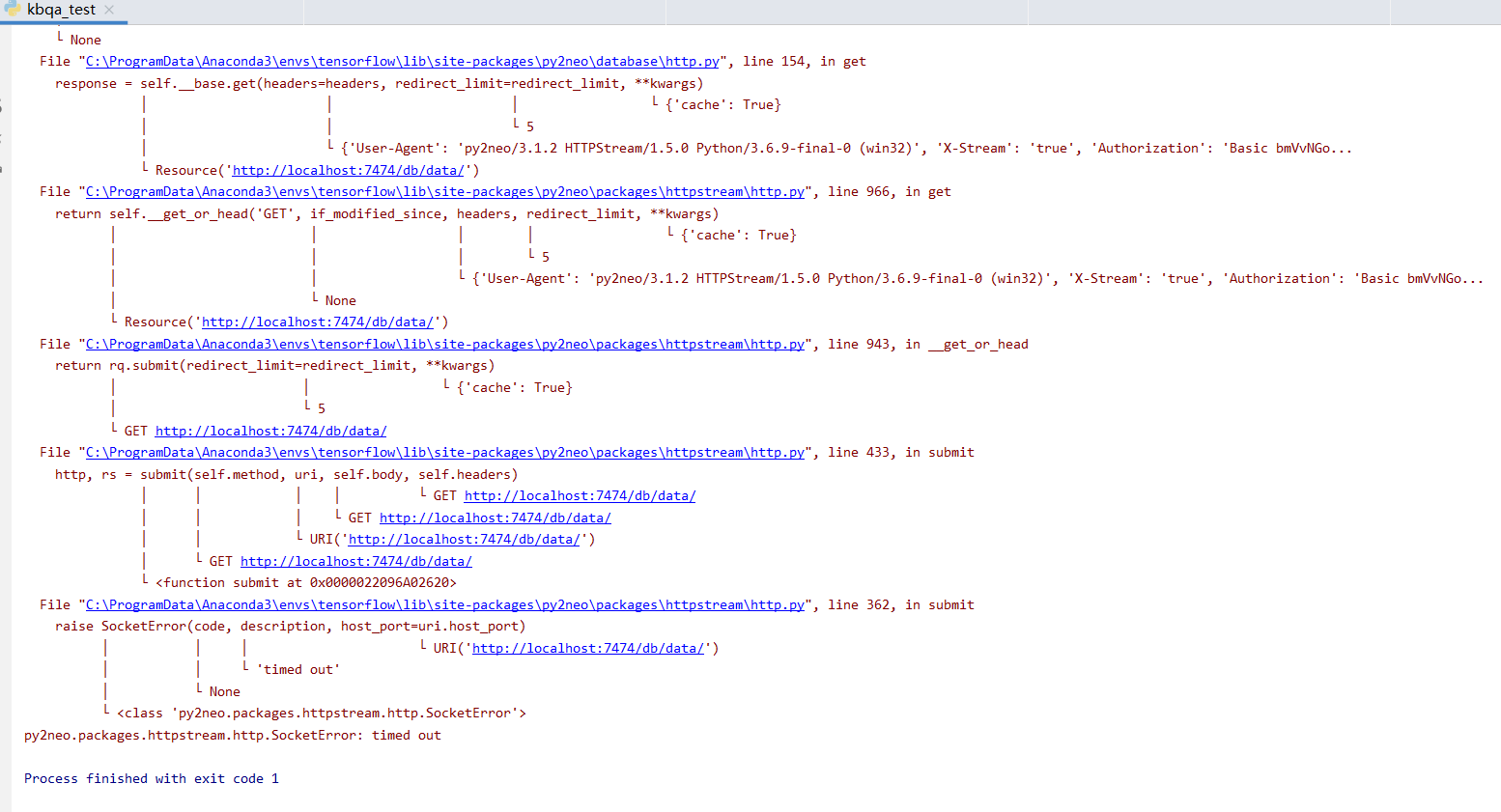
将py2neo从2020.1.1降级到3.1.2解决

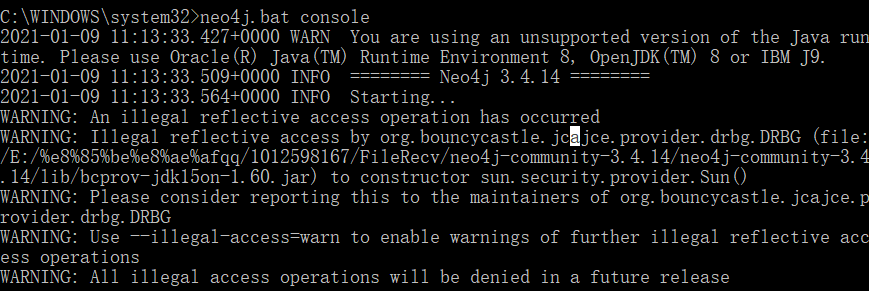


切换到管理员的cmd解决。

运行继续报错

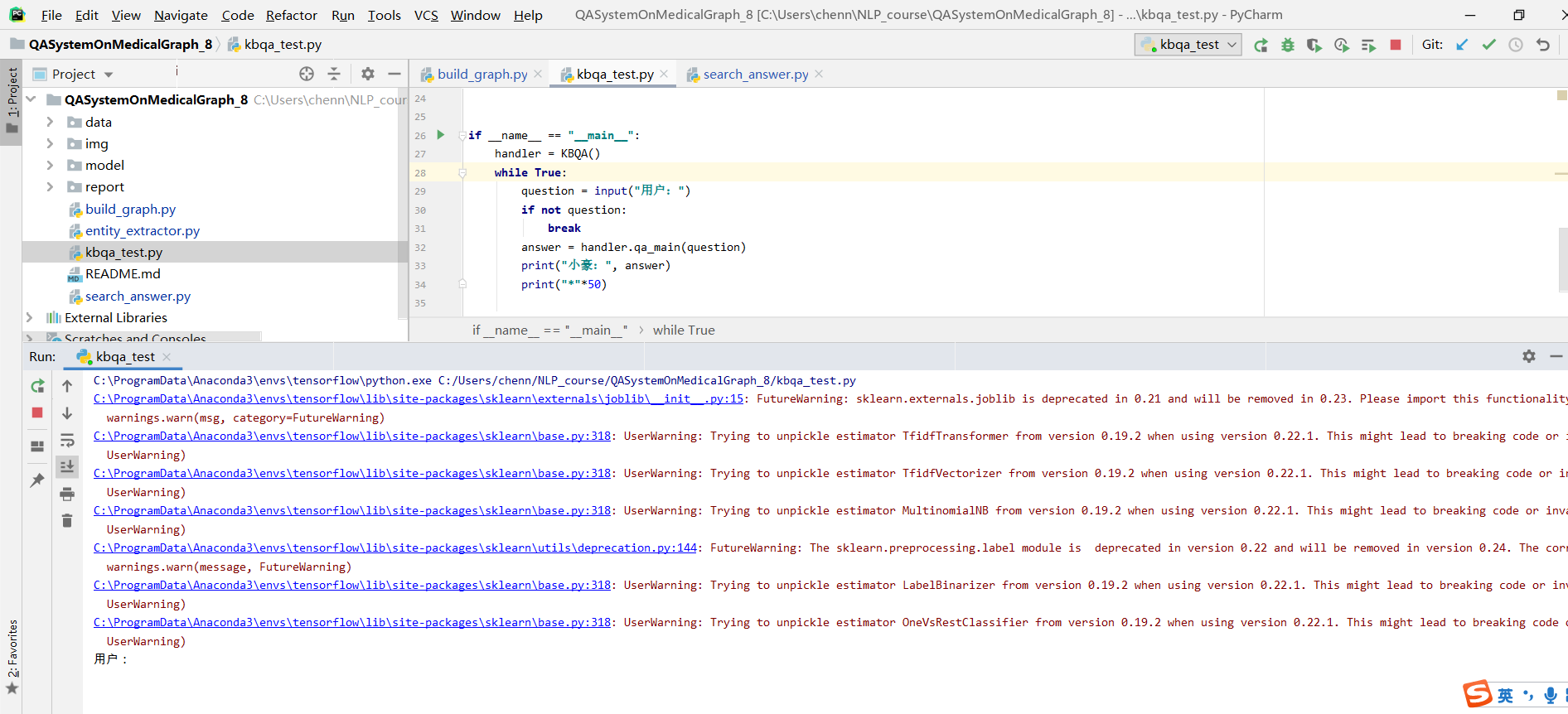


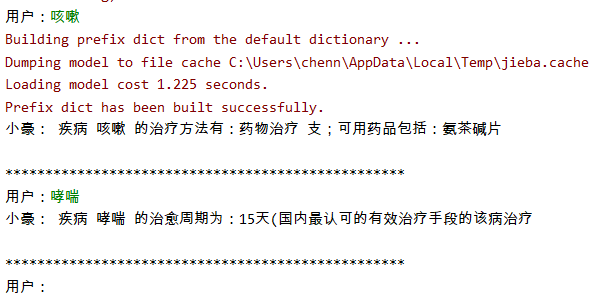




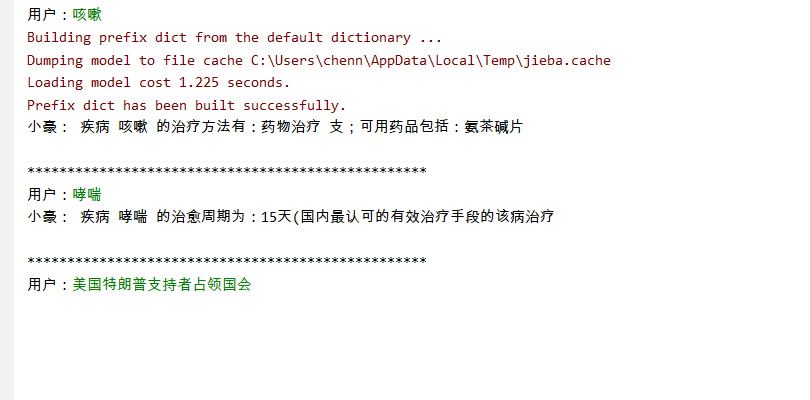
直接重启

最后 效果如下





输入错误则会卡住



5.3运行环境：注意py2neo3.1.2

Python3 数据库：neo4j 预训练词向量：

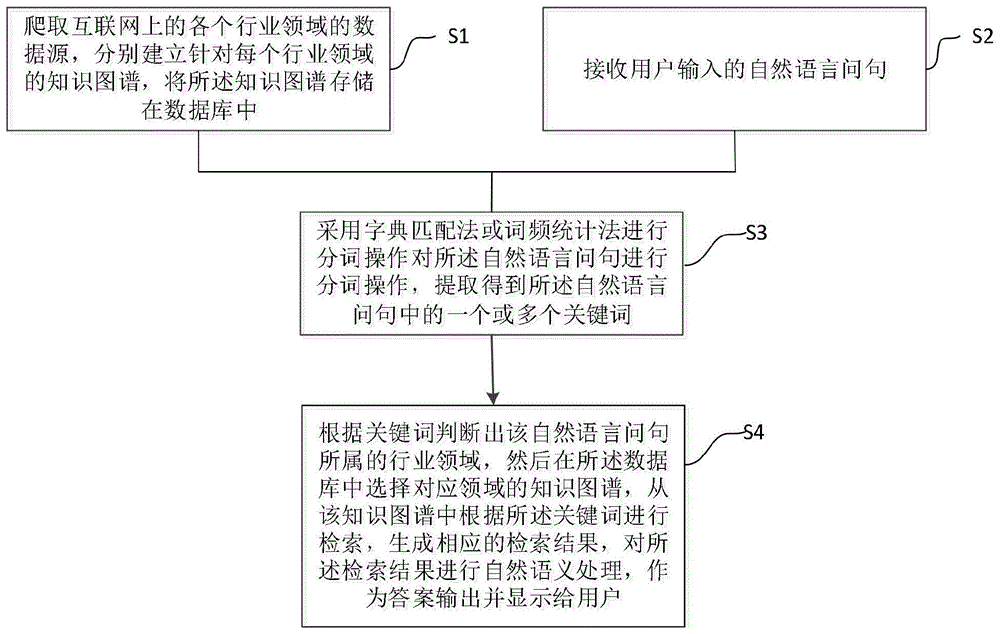
<https://github.com/Embedding/Chinese-Word-Vectors或><https://pan.baidu.com/s/14JP1gD7hcmsWdSpTvA3vKA>

5.4 参考链接

https://github.com/zhihao-chen/QASystemOnMedicalGraph

**六、实验感想**

知识图谱项目的整个流程大体如下



提出一种基于知识图谱的智能问答方法及系统

包括如下步骤：

步骤s1，爬取互联网上的各个行业领域的数据源，分别建立针对每个行业领域的知识图谱，将所述知识图谱存储在数据库中；

步骤s2，接收用户输入的自然语言问句；

步骤s3，采用字典匹配法或词频统计法进行分词操作对所述自然语言问句进行分词操作，提取得到所述自然语言问句中的一个或多个关键词；

步骤s4，根据步骤s3中得到所述关键词判断出该自然语言问句所属的行业领域，然后在所述数据库中选择对应领域的知识图谱，从该知识图谱中根据所述关键词进行检索，生成相应的检索结果，对所述检索结果进行自然语义处理，作为答案输出并显示给用户。

进一步，在所述步骤s1中，建立知识图谱，包括如下步骤：对爬取得到的互联网上的各个行业领域的数据源，提取每项数据的标题和正文，然后进行语句拆分，并根据预设的过滤原则对拆分结果进行过滤，以去除冗余的词语，保留关键信息文本，根据所述保留的关键信息文本进行语义分析，得到该数据源所属领域，并在所述数据库中对该数据源的领域进行标记，以便后续查询使用。

进一步，在所述步骤s2中，所述用户输入的自然语言问句为文本形式或语音形式，当所述自然语言问句为语言形式时，将所述语音进行文字识别处理，生成对应的文本语言。

进一步，在所述步骤s4中，所述从该知识图谱中根据所述关键词进行检索，生成相应的检索结果，包括如下步骤：

根据多个关键词在所述知识图谱中进行检索，查询得到多项检索结果；

将所述多项检索结果进行相关性排序，将与所述关键词关联度最高的检索结果，再进行自然语言整理后，作为答案输出给用户查看。

进一步，在所述步骤s4中，在将答案输出给用户查看的同时，在输出界面的下方向所述用户同步推送所述自然语言问句的相关领域信息。

本发明实施例还提供一种基于知识图谱的智能问答系统，包括：知识图谱建立模块，用于爬取互联网上的各个行业领域的数据源，分别建立针对每个行业领域的知识图谱，将所述知识图谱存储在数据库中；数据库，用于存储所述知识图谱；问句输入模块，用于接收用户输入的自然语言问句；分词模块，用于采用字典匹配法或词频统计法进行分词操作对所述自然语言问句进行分词操作，提取得到所述自然语言问句中的一个或多个关键词；检索模块，用于根据所述关键词判断出该自然语言问句所属的行业领域，然后在所述数据库中选择对应领域的知识图谱，从该知识图谱中根据所述关键词进行检索，生成相应的检索结果；自然语义处理模块，用于对所述检索结果进行自然语义处理；答案输出及显示模块，用于将自然语义处理后的答案输出并显示给用户。

进一步，所述知识图谱建立模块对爬取得到的互联网上的各个行业领域的数据源，提取每项数据的标题和正文，然后进行语句拆分，并根据预设的过滤原则对拆分结果进行过滤，以去除冗余的词语，保留关键信息文本，根据所述保留的关键信息文本进行语义分析，得到该数据源所属领域，并在所述数据库中对该数据源的领域进行标记，以便后续查询使用。

进一步，所述用户输入的自然语言问句为文本形式或语音形式，当所述自然语言问句为语言形式时，由所述问句输入模块将所述语音进行文字识别处理，生成对应的文本语言。

进一步，所述检索模块根据多个关键词在所述知识图谱中进行检索，查询得到多项检索结果；将所述多项检索结果进行相关性排序，将与所述关键词关联度最高的检索结果，再进行自然语言整理后，作为答案输出给用户查看。

进一步，所述答案输出及显示模块在将答案输出给用户查看的同时，在输出界面的下方向所述用户同步推送所述自然语言问句的相关领域信息。