**《知识工程》课程报告**

姓名

学号

专业

班级

天津大学智能与计算学部

2020年 5 月

1.完成数据准备

为了获取疫情相关的实时数据，保持数据的准确性并能够快速更新问答系统需要借助python的网络爬虫功能进行相应的数据获取

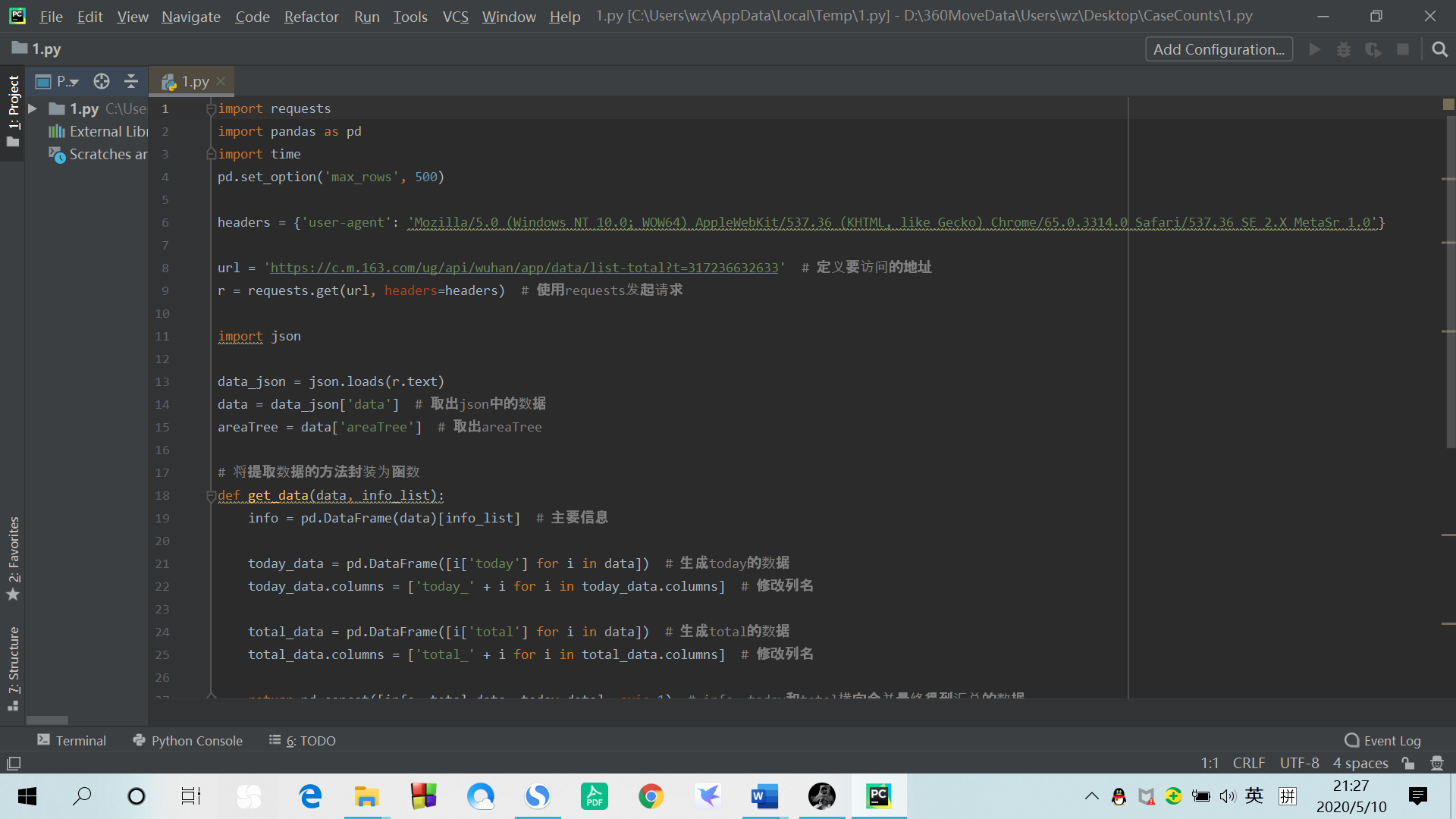
使用python的爬虫进行网络疫情数据的获取，获取数据的网站为网易提供的全球实时疫情动态网站

<https://wp.m.163.com/163/page/news/virus_report/index.html?_nw_=1&_anw_=1>



编写代码的时候需要了解有关网页方面的知识，并要求获取的数据具有规范的格式（如.csv）

具体的python爬虫代码部分截图



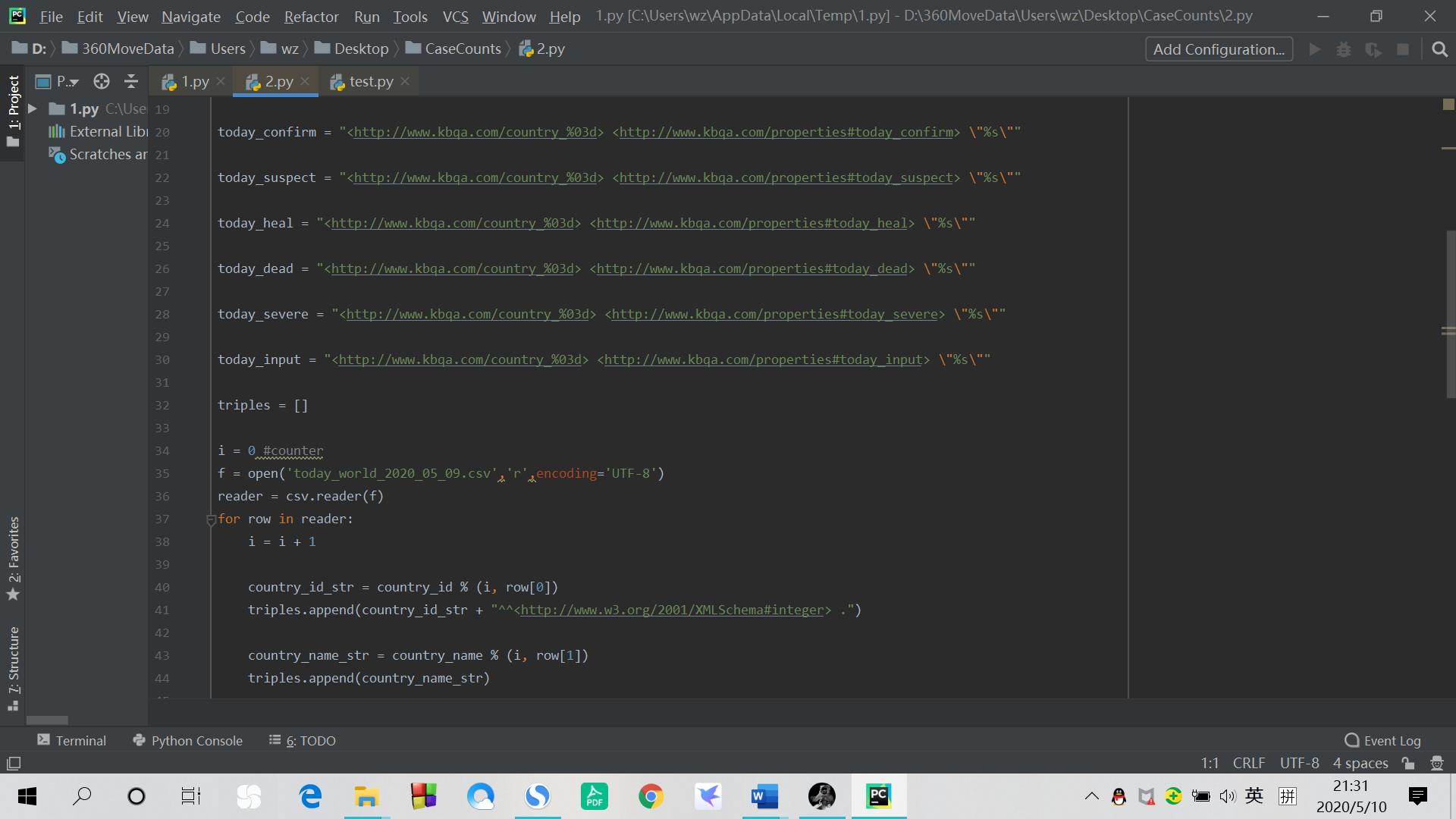
获取到的数据为格式为：

各国的id，名称，总感染数，总疑似感染数，总死亡数，总治愈数，总危重症病人数，总输入病例数，新增感染数，新增疑似病例数，新增死亡数，新增危重症病人数，新增输入病例数。具体数据在 today\_world\_2020\_05\_09.csv文件 中

2.构建RDF知识库

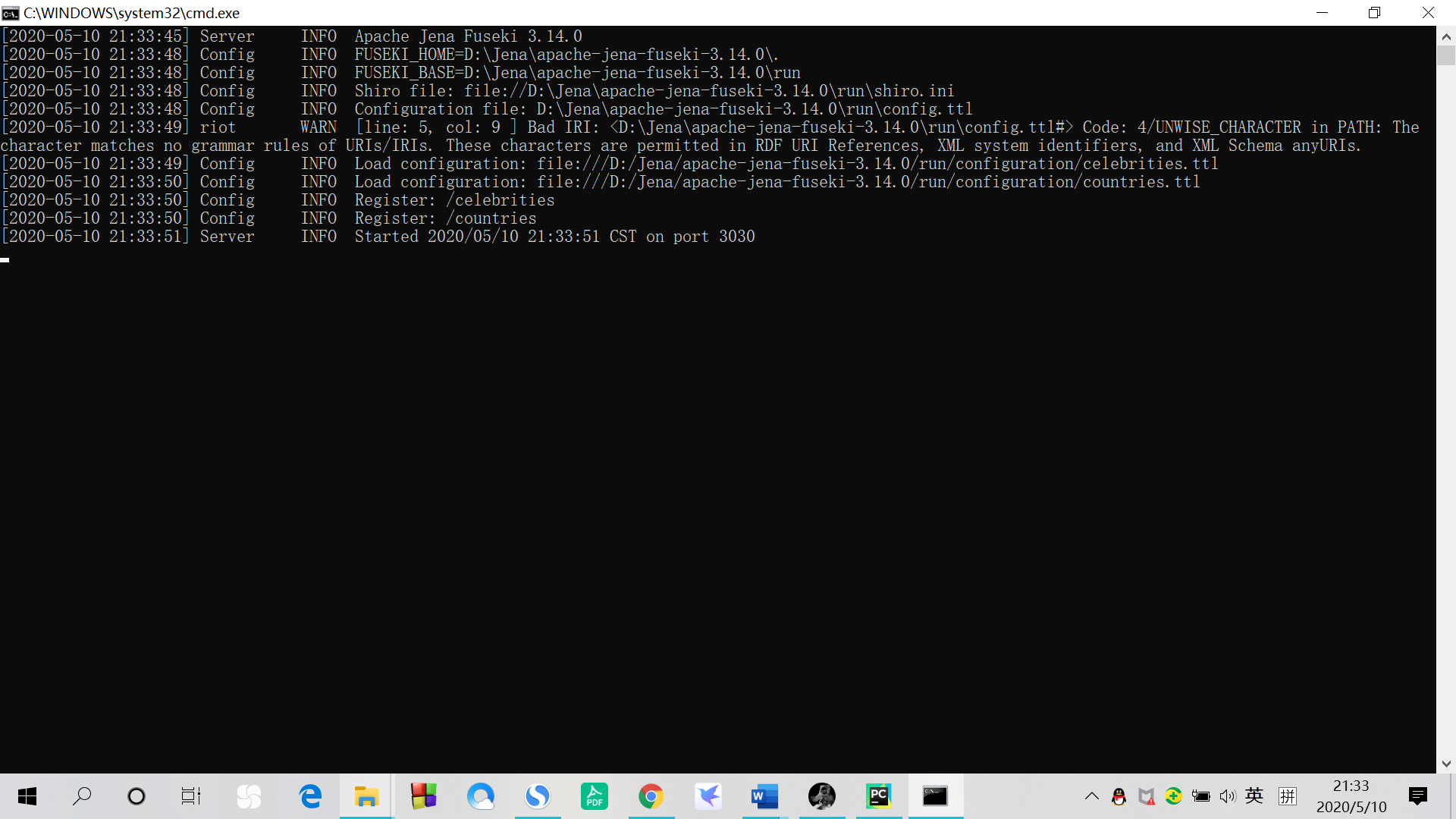
需要了解三元组文件的具体格式和语法（例如，如何使数字从字符串格式转换为可以进行大小比较的整型格式）

使用python的字符串操作进行nt文件的转换，具体代码的部分截图

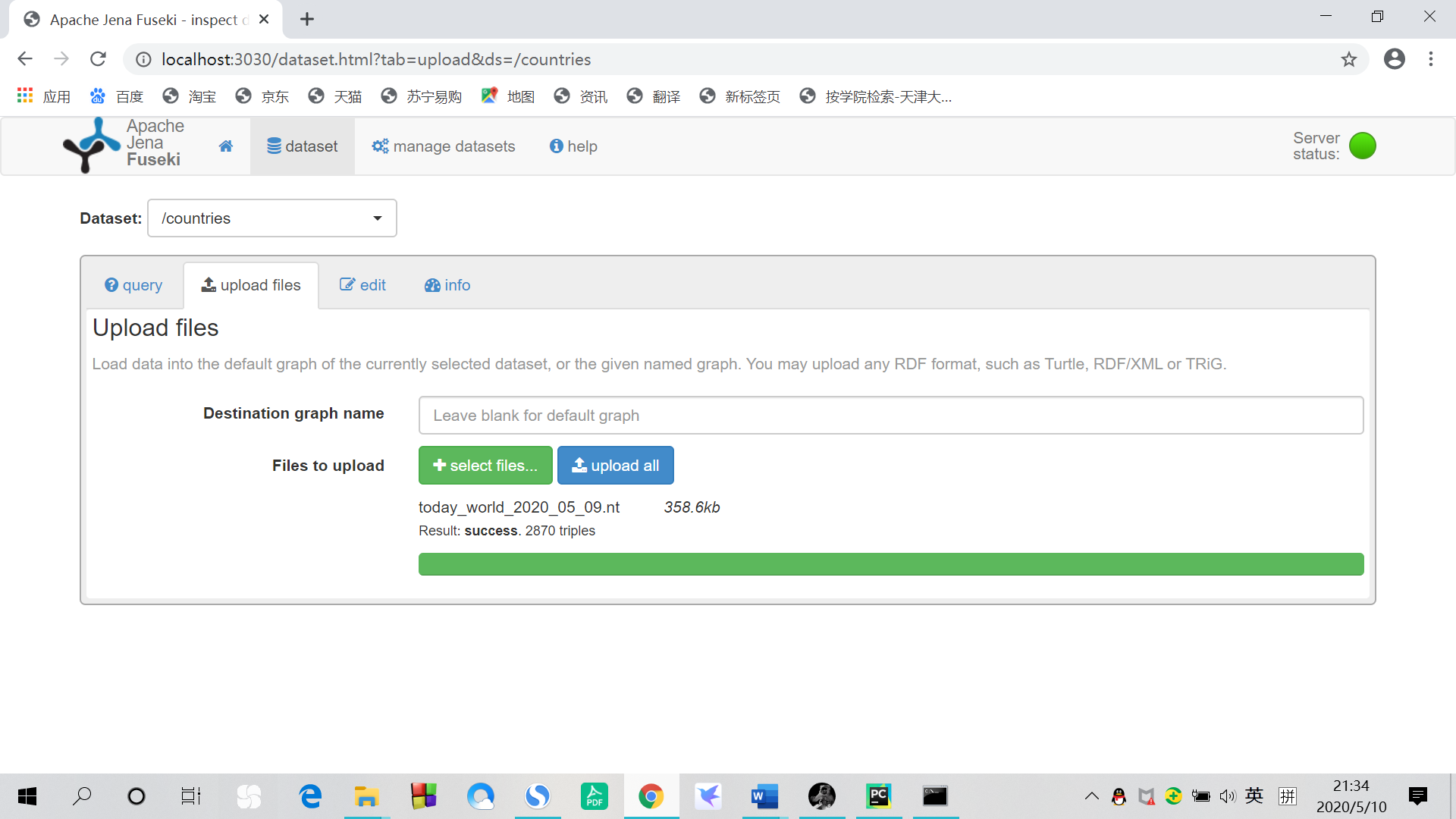


之后导入到jena中进行查询（需要注意三元组文件的相关语法，否则导入时会报错）

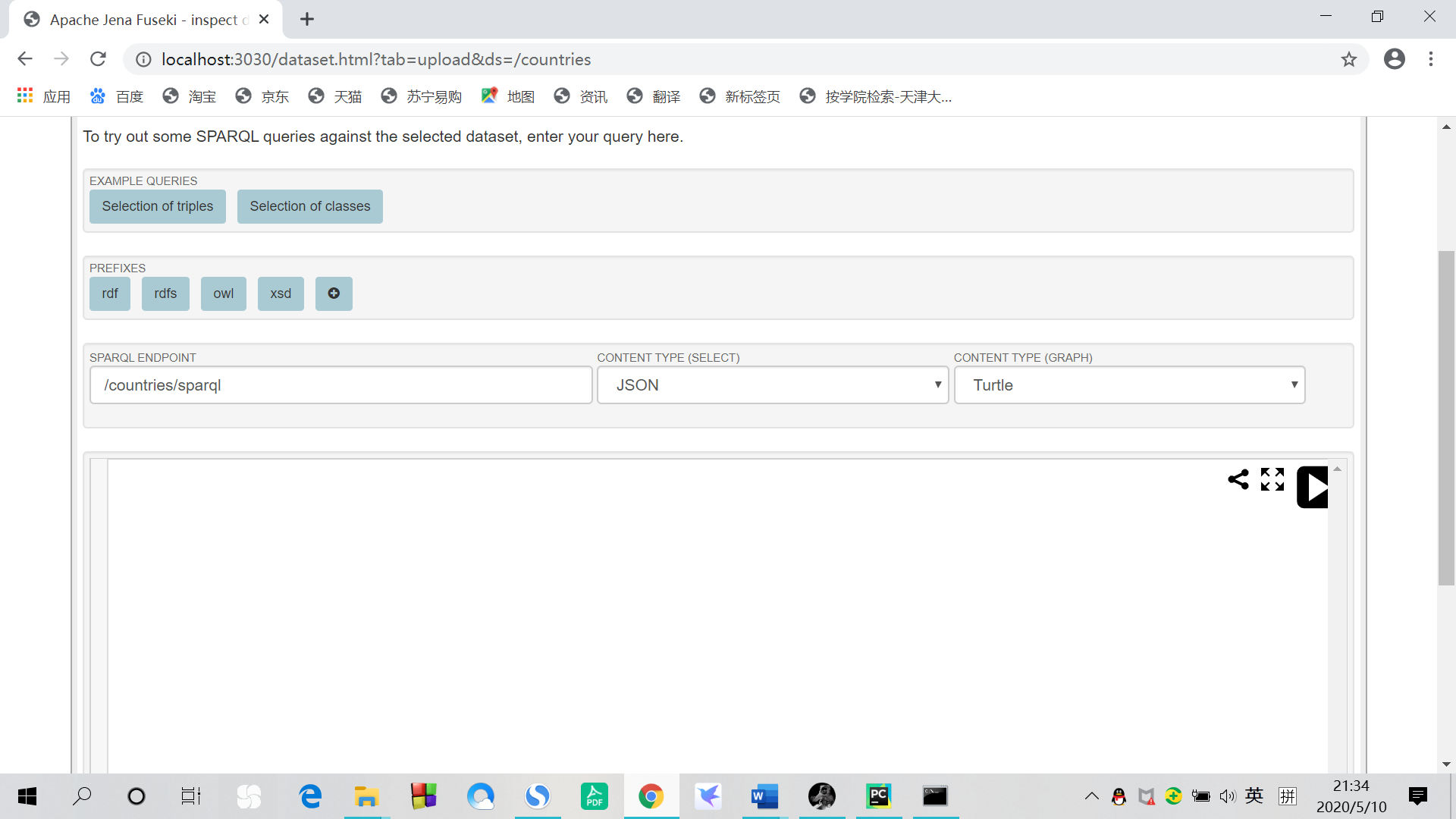
打开Jena



导入数据



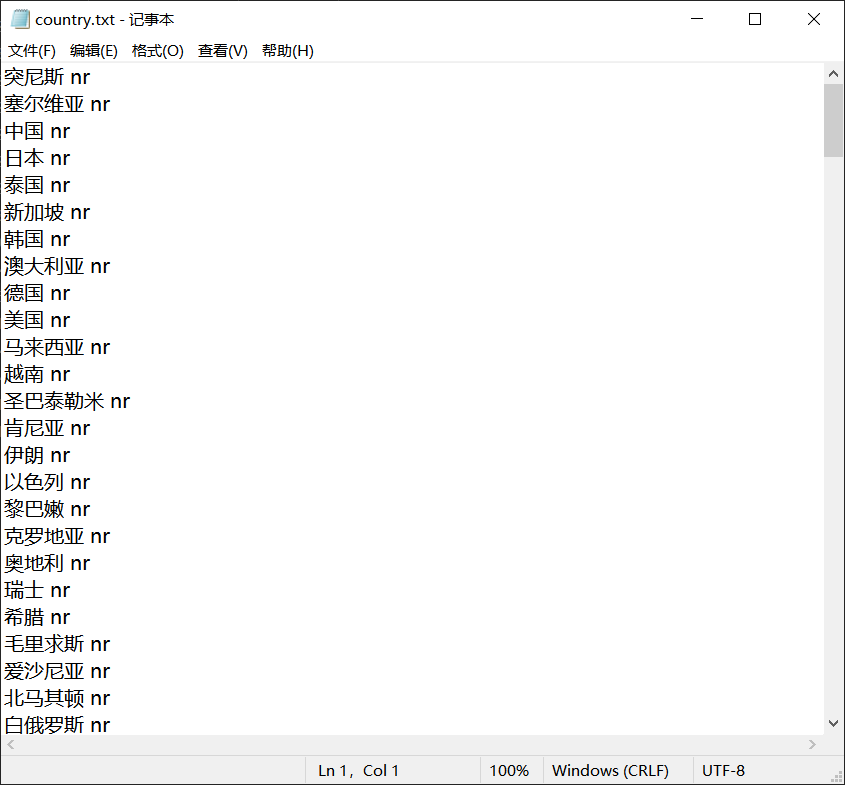
查询界面



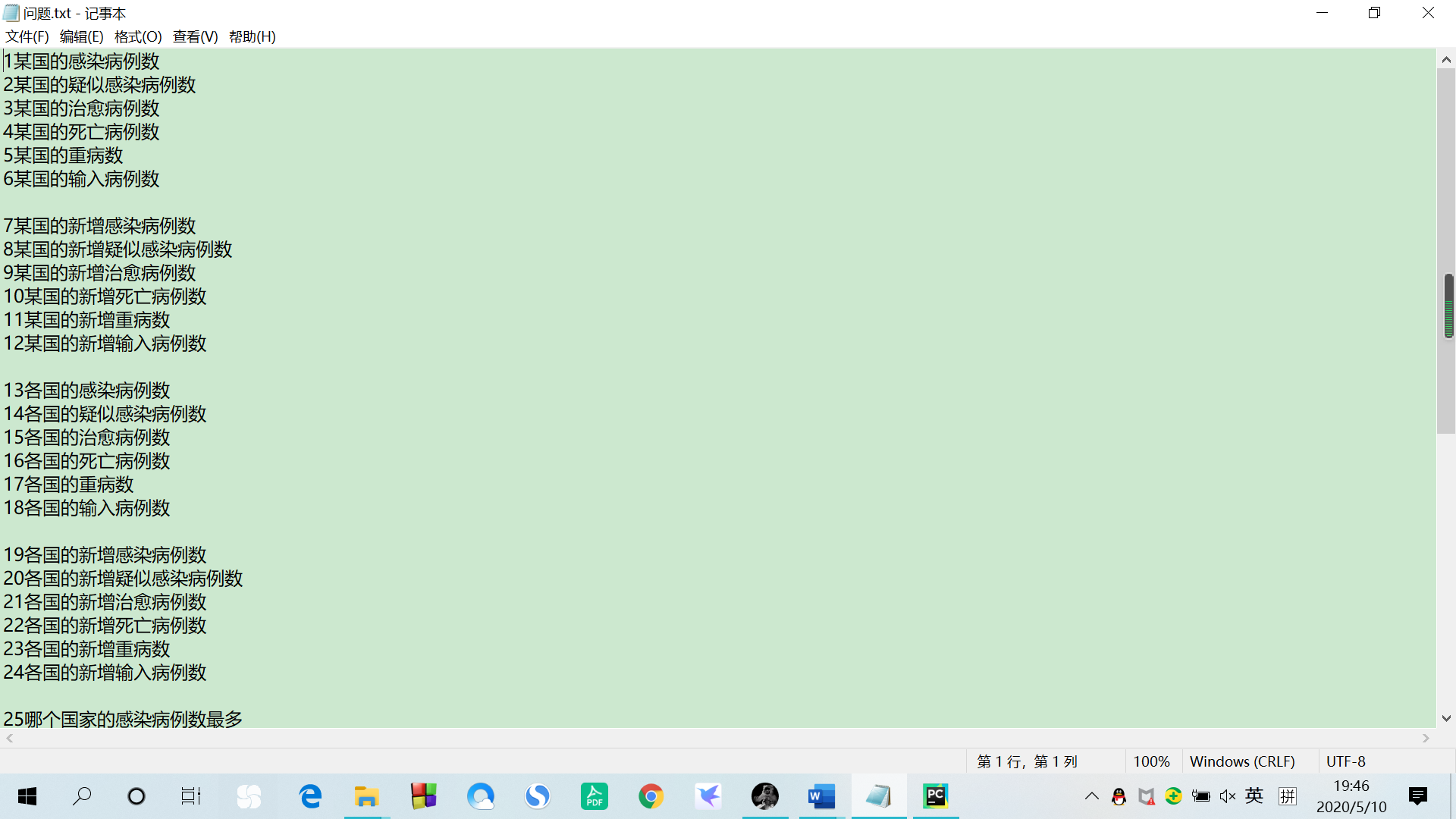
3.根据知识库设计模板问题并编写SPARQL查询模板

本系统的问答系统部分是通过正则匹配加sparql语句查询实现的。通过jieba中文分词模块将输入的问句按照词语进行切分，并对切分得到的词语进行词性标注，由于jieba分词模块对于国家的名词的切分不满足可扩展性的需要，所以我们为模块添加了外部字典来辅助分词。

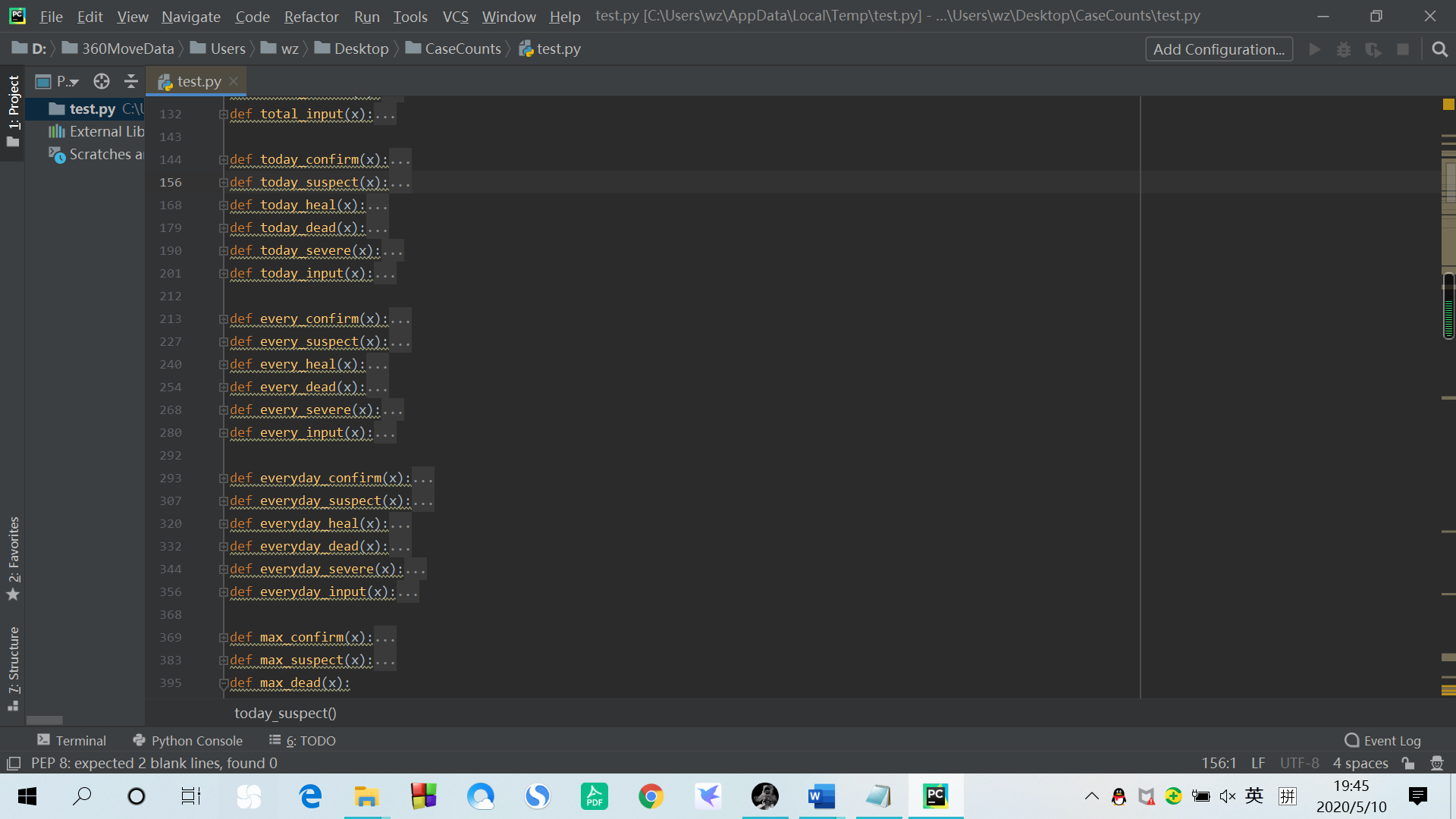
外部词典



问答系统的问题模板



问答系统函数实现的代码



（具体编写过程中需要考虑到各种查询的需求，不断地丰富模板的内容）

4.实现问题的正则匹配

在正则匹配部分，最重要的就是关键词设置，通过设置不同的关键词来匹配不同的问题，我们在问答系统中针对预设的问题编写了相应的正则匹配规则。这样，向系统输入问题时，就能通过对问题分词结果的正则匹配找到用户想问的问题，并能生成相应的sparql查询语句，在nt格式的数据中查询获得结果。对问答系统我们预设了以下关于全球各个国家疫情感染情况的规则。

（需要注意有的语句在匹配的时候可能匹配到多个问题，需要把较强的条件放到前面，以便在匹配的时候不会出现同时匹配两个或多个规则的情况。）



5.可视化

问答系统的一个特色功能是能够根据用户输入的问题，打开mp4视频格式的可视化图谱，例如用户想要查询某个国家的疫情死亡病例最多，问答系统在输出结果的同时，会打开一个相应的mp4格式的视频，以可视化的形式展示了这个国家的死亡病例数量，能够更加直观地观察到该国死亡病例数量和其他国家的死亡病例数量的对比情况。例如当用户想要了解全球各个国家的疫情情况时候，系统就能够生成一个全球环视图，同时用三维柱状模型代表当前用户想要了解的数据，并以视频的形式展现出来，用户可以通过视频直观地了解到各个国家之间染疫情况的差别。

（利用Power Map软件进行视频的生成，增加了渲染和视频的不同播放效果）



可视化方面我们还生成了累计确诊数，治愈数，死亡数等的世界地图，可以通过颜色的深浅和鼠标选中直观的看到世界疫情的集中度和各国的相关人数。

