

# 中间件课程报告

张杜璠

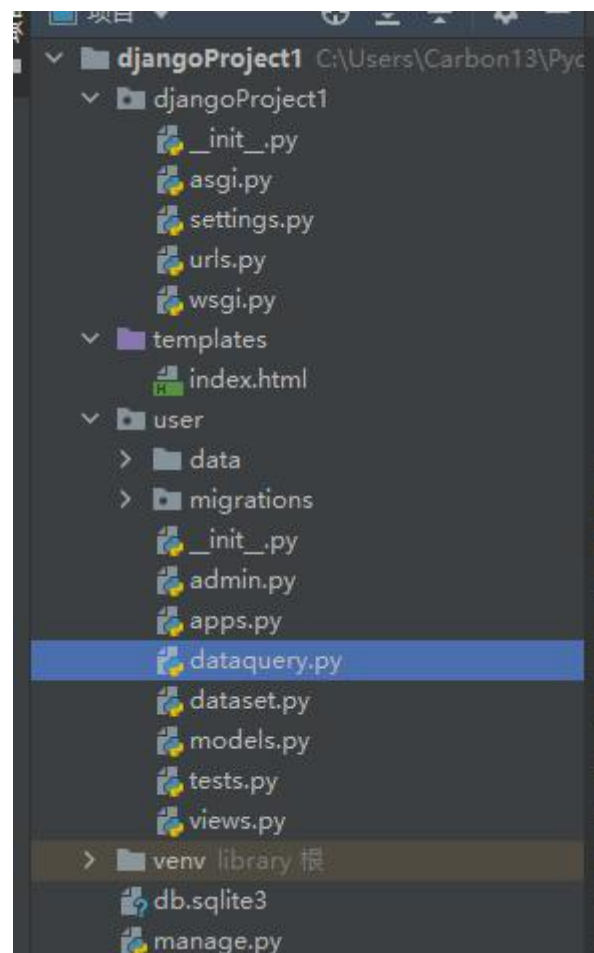
2020104256

选择任务：task2

项目背景：知识图谱（Knowledge Graph），在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

知识图谱，是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。

项目结构：



django 结构，data 存放数据集和数据清洗结果

user 是自建 django 的 app

dataquery 存放数据处理相关代码

view 存放界面

使用技术：python, numpy, pandas, PyPDF2, textract, py2neo, bibtexparser, pdfminer, sklearn, neo4j, django

项目概述：使用课程提供的数据集完成，历时三晚上（上班）从零开始一边学习一边做。实现的功能搜索作者获取作者论文，同时推荐相关领域文章。同理也可从作品和领域进行查询。

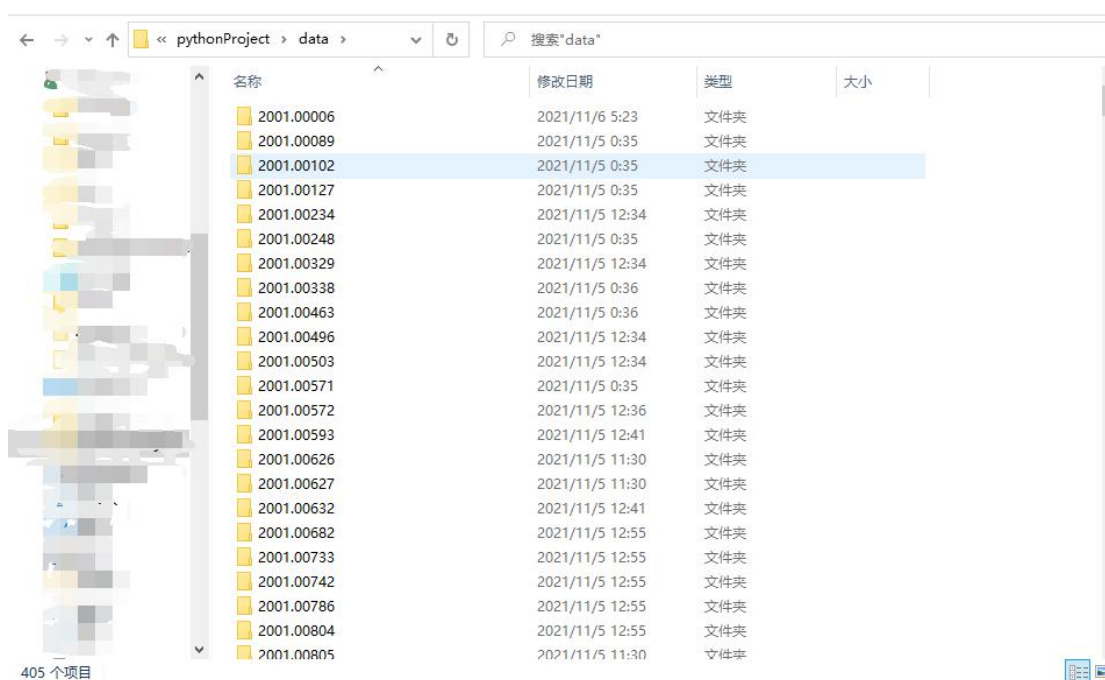
```
C:\Users\Carbon13\PycharmProjects\pytho
查询模式（1.查作者，2.查文献，3.查领域）：
```

查询模式

内容

代码主要包含数据处理，图谱生成，交互三部分。

数据处理上说实话没有研究透数据集，对 pdf 和 bib 的处理有些粗糙，脏数据的清洗还不够全面。原数据集使用了六分之一左右（磁盘不够了）



针对 pdf 先后尝试 PyPDF2, textract, pdfminer 进行处理, bib 原本使用自己开发的代码后来感觉效果较差选择使用 bibtexparser 但其很久未更新有较大问题, 在自己代码上修改使其兼容, 总之数据处理占了绝大多数时间。到头来也不太会看 bibtex 文件。

数据库使用 neo4j, 采用原生 cql 进行操作。使用那个 django 做了简单界面

结果:

```
查询模式（1.查作者，2.查文献，3.查领域）：1
查询的作者名字：Murphy, Kevin P.
该作者的作品：
['Machine Learning: a Probabilistic Perspective']
该作者的领域：
['{UCI Machine Learning Repository}', 'Multiagent Systems: A Survey from a Machine Learning Perspective']
查询模式（1.查作者，2.查文献，3.查领域）：|
```