****

**人工智能程序设计(python)**

**实验报告**

**实验题目： 设计一个程序抽取百度百科人物知识**

**专 业： 人工智能**

**姓 名： 马承乾**

**实验日期**：  **2023.3.23**

**指导教师： 曹冬林**

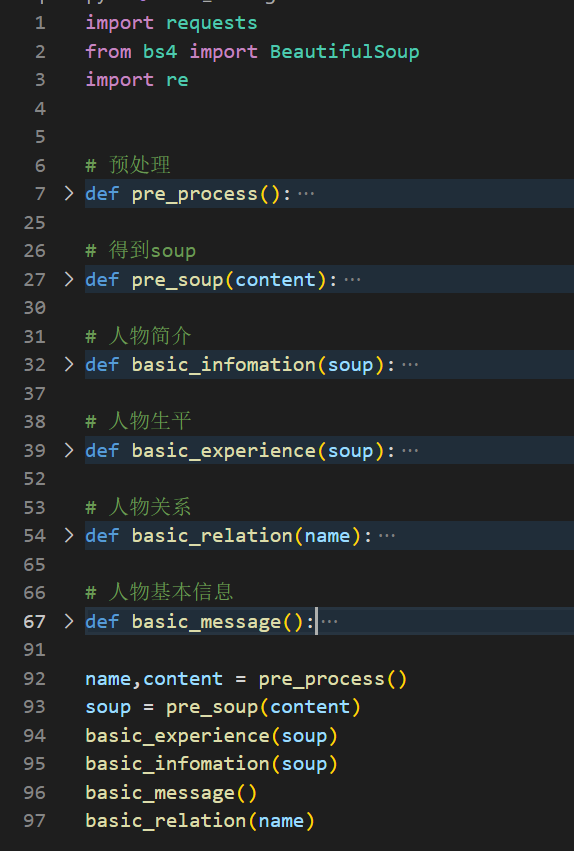
1. **实验目的**

设计一个程序抽取百度百科人物知识。

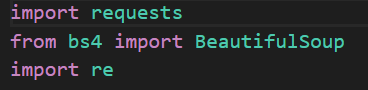
1. **实验内容**

抽取简介，基本信息，人物履历，关联人物。

1. **实验步骤以及结果**
2. 代码总览：



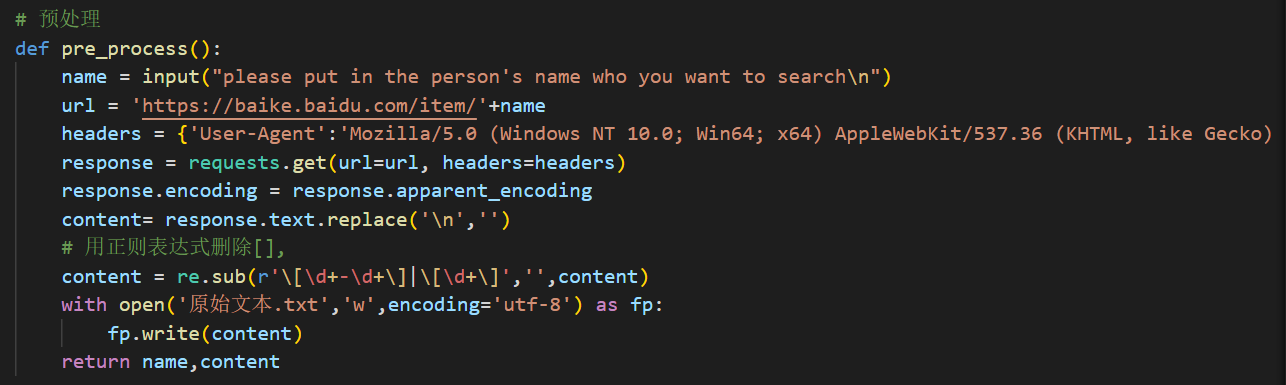
1. 细节分析：引入需要的库：

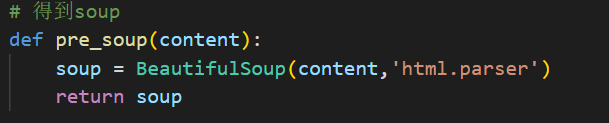


1. 进行预处理，得到人物姓名，网页中的html格式内容和对应的网页中的soup对象；



1. 获得人物姓名，用requests获得网页内容，使用正则表达式删去对人物的引用信息；删除网页内容中的’\n’符号，随后写入原始文本，用以对照；

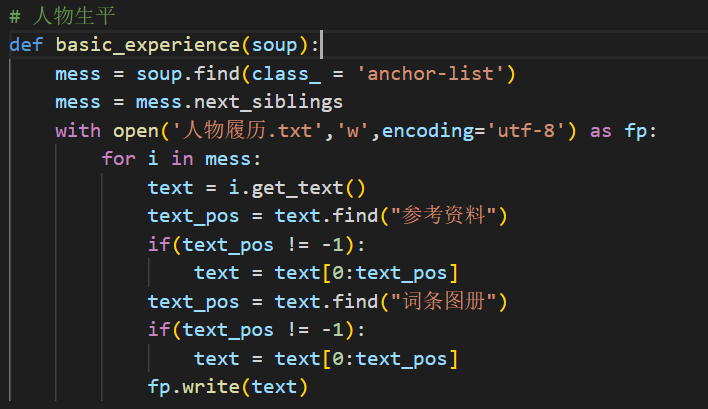
通过bs4库获得对应的beautifulsoup对象



1. 通过soup获得人物履历：



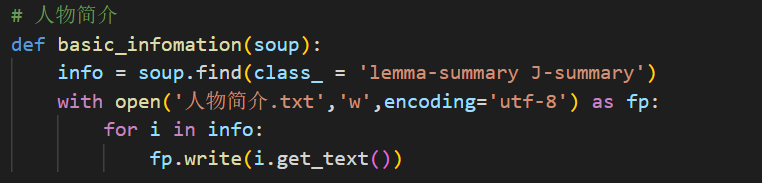
注意到人物履历之前有个标签class\_属性为anchor-list,找到这个tag，其后的内容用next\_siblings访问，通过get\_text()方法得到tag中的string类型内容，若内容中包含参考资料，词条图形内容则用string方法直接删去。



1. 通过soup得到基础信息



注意到基础信息对应的标签的class\_信息都是lemma-summary J-summary，所以可以从此提取基础信息



1. 通过soup获得人物基本信息；



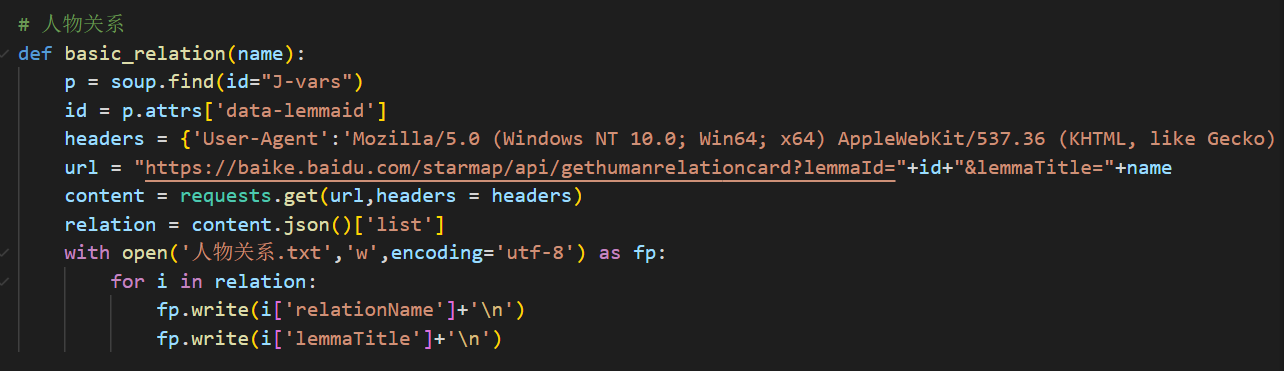
通过正则表达式获得人物信息，随后使用正则表达式进行数据清洗；



1. 获得人物关系

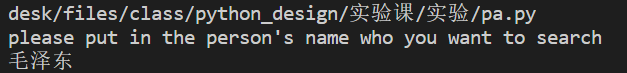


注意到有关人物均放置在id为J-vars的tag中，这个属性其后的就是我们所查找的人物在百度百科对应的号码，对应着属性data-lemmaid，获取data-lemmaid之后即可组成新的链接，该链接的json内容就是我们要的相关人物。

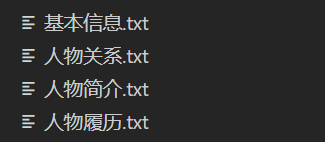


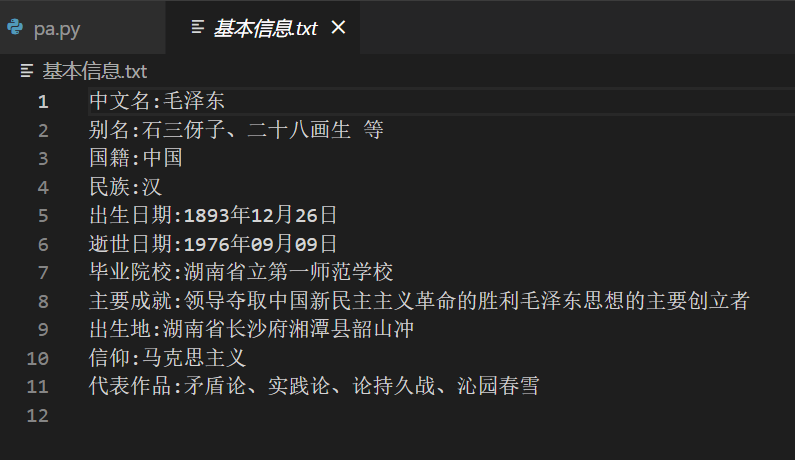
1. **实验结果与分析**

运行过程：



以下为获取的信息储存情况：

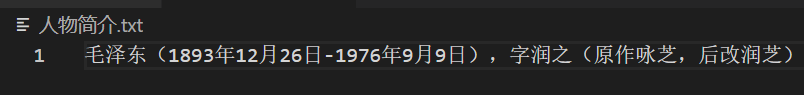


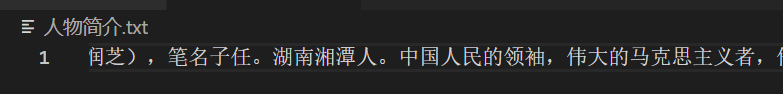
以下为基本信息：

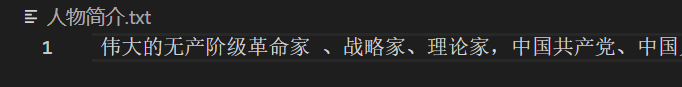
以下为人物关系（因较长，故只在此展示部分，详情可使用代码自行实验）：



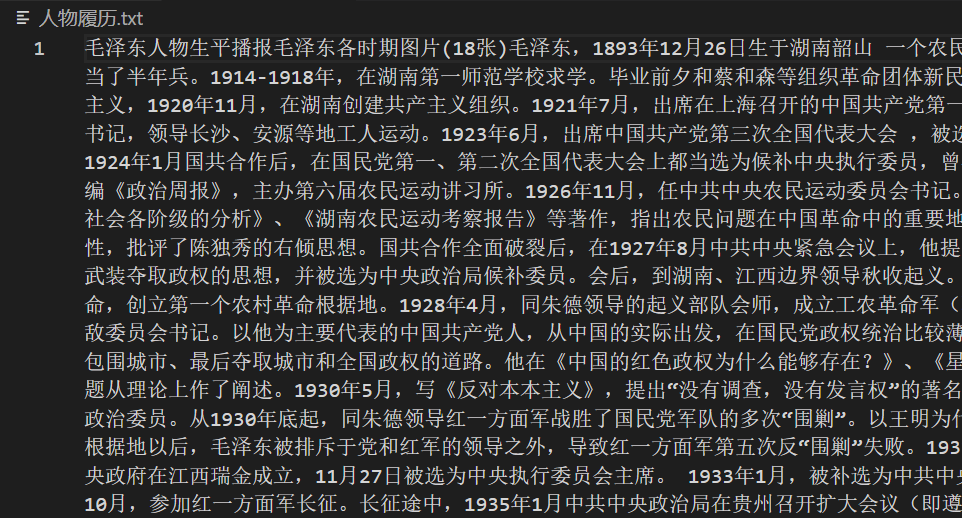
以下为人物简介（因较长，故只在此展示部分，详情可使用代码自行实验）：







以下为人物履历（因较长，故只在此展示部分，详情可使用代码自行实验）：



1. **实验总结**

通过本次实验，我学会了使用python语言中的诸多第三方库，如beautifulsoup解析html文件，用requests获取网页的html文件，用re库对获取到的html文件做数据清洗，用字符串方法对html内容做删除和保留，极大地提升了我的python水平和工程能力，以及对网页结构的理解。