任务教程有以下部分内容：

1．四领域（娱乐、政治、历史、体育）人物更细致的分类

2．具体分类需要抽取的详细属性

3．本科生任务：

如何考虑多种情况通过正则对摘要进行抽取

如何验证所写正则表达式正确性以及将抽取好的数据进行拼接

本科生要做哪些属性的模板

1、2 部分为了解内容，3 是具体的任务教程和任务分配

schema 模板

在百度百科人物分类的基础上，对部分重合项进行删改，最后得到划分后更细致的分类：

娱乐明星

域内更细致分类：歌手 演员 导演 模特 主持人 编剧

人物属性值：中文名 别名 国籍 民族 籍贯 出生日期 职业 主要成就

人物关系：父亲 母亲 丈夫 妻子 儿子 女儿 兄弟 男友 女友 姐妹 好友 同学 老师 同事 经纪人

政治人物

域内更细致分类：国家元首 军事人物 政府官员

人物属性值：中文名 国籍 民族 籍贯 出生日期 毕业院校 任职

人物关系：父亲 母亲 丈夫 妻子 儿子 女儿 兄弟 姐妹 好友 同学 老师 同事 前任 后任 直接领导 直接下属

历史人物

域内更细致分类：年代划分 三皇五帝夏商周，春秋战国秦两汉，三国两晋南北朝，隋唐五代和十国，辽宋夏金元明清。

人物属性值：本名 别称 字号 民族 所处时代 籍贯 出生日期 去世日期 重要事件/主要作品/主要成就

人物关系：父亲 母亲 丈夫 妻子 儿子 女儿 兄弟 姐妹 好友 同学 老师 幕僚 君臣

体育人物

域内细致分类：篮球运动员 足球运动员 棋手 教练 乒乓球运动员

人物属性值：中文名 外文名 别名 国籍 民族 籍贯 出生日期 身高 运动项目 主要奖项

人物关系：父亲 母亲 丈夫 妻子 儿子 女儿 兄弟 姐妹 好友 同学 老师 同事 教练

正则表达式抽取

参考资料

正则表达式规则：https://www.runoob.com/regexp/regexp-syntax.html

正则表达式验证工具：https://tool.oschina.net/regex/

思路：

分析某一类型属性的特征，提取主要特征，用正则表达式进行提取。写正则时要考虑对于一个百科人物的摘要该怎么设计模板才能避免一些问题，下面在举例中讲解一下。

举例：例如从以下文本中针对出生年月日进行提取

```
str1 ='<http://zhishi.me/baidubaike/resource/许耀元>
<http://zhishi.me/ontology/abstract> "许耀元，男，汉族，1952 年 12
月生，江苏吴江人。1971 年 10 月加入中国共产党，1969 年 2 月参军入伍."
```

可以看出模式为 1952 年 12 月出生，同时结合其他样例，可以得到“日期(必须包含年份，月份和日期可省略)[(出)生/-]（出字也可以省略)”的基本模板，这里如果简单的写出来“XX 年 XX 月”的模板识别，会识别出来多个（包括出生时间、入党时间、入伍时间)，这样就无法有针对性的对生日进行补全，所以我们考虑摘要中陈述一个人的生日可能出现的情况“XX 年 XX 月生”、“XX 年生”“生于 XX 年”等等。同时也要想到“XX 年生人”和“XX 年生”只要写出来后者的模板就可以囊括两种情况啦；以及“XX 年生”和“XX 年 XX 月生”、“XX 年 XX 月 XX 日生”也可以写为同一个模板，只要将“月”和“日”的部分设为可选项即可，这些内容在学习正则表达式后就会明白。再举个例子，如果我们要去写“籍贯”的模板，不可能只有“籍贯为 XX”一种表达形式，还有可能“出生于（中文字符)”这样，所以要考虑多种表达习惯。最后再将模板转化为正则表达式，这一步举个简单的例子：

```
regex1 = r'(\u4e00-
\u9fa5)*(\d{1,4})+[\u5e74.-]+(\d{1,2})*[\u6708.-]*(\d{1,2})*[\u65e5.-]*((
\u51FA*\u751F+)|-)'
#匹配模式 1952 年 12 月[(出)生/-]
```

其中'\u4e00-\u9fa5'表示中文字符，\u5e74 表示中文字符年的 Unicode 编码，\u6708 表示中文字符月的 Unicode 编码，\u65e5 表示中文字符日的 Unicode 编码

得到的匹配结果为[('', '1952', '12', '', '生', '生')]，后续会将年份月份进行拼接为 1952-12-(空格) (归一化后的属性样式) 的形式补全到属性里，得到需要的结果。

针对其余的情况，我们可以采用不同的正则表达式进行应对

```
regex2 = r'(\u751F\u4E8E)(\u4e00-
\u9fa5*)(\d{1,4})+[\u5e74.-]+(\d{1,2})*[\u6708.-]*(\d{1,2})*'
#匹配模式 生于 1952 年 12 月
```

验证工具验证

可以直接使用 python 编程进行验证，也可以使用在线工具进行调试，示意图如下：

归一化要求

针对目前的属性值，仅对时间类别，书名类和数字类属性有归一化要求，时间类的归一化要求为 1992-09-09，如遇到只有年份没有具体月份或日期，则对应位置留空，如 1992-空格-空格，数值类的归一化要求为数值加中文数字单位，如 182 厘米，70 公斤，主要作品需要添加书名号《》，如《红楼梦》

任务流程：

（1）根据数据文件，查找并分析一种属性通常有哪些表示方法，如身高属性可以在数据文件中找出“1.77 米”“175cm”“184 厘米”“170.0cm”“1.85m”“186”“172 公分”“1 米 85”“1.96 米/6 英尺 5 英寸”等等表示方法。

（2）分析哪几种表示方法可以共用一个模板表示出来

（3）将模板转化为正则表达式，运用验证工具进行准确性验证

（4）提取匹配的结果，拼接成归一化后的结果

同学们先用发的旧的百度百科数据作为参考，后续会进行更新

任务内容-需要写模板的属性：

① 娱乐明星：出生日期 职业 主要成就

② 政治人物：出生日期 毕业院校 任职

③ 历史人物：出生日期 去世日期 重要事件/主要作品/主要成就

④ 体育人物：出生日期 身高 运动项目 主要奖项

先暂时安排这些大家上手练一练，所有人都有的出生日期方便上手，如果有更好的想法，想给自己负责的任务领域加上一些更具有普遍性和代表性的属性/关系做一做也非常欢迎大家主动提出来，后续可能有新的任务，辛苦啦