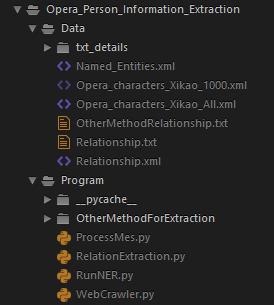
网络爬虫与信息抽取

王瀛 2017213846

**1.提交文件内容**

提交文件夹包含两部分内容，如下图所示：



其中第一部分是Data文件夹，其中包含txt\_details文件夹，里面是1000个京剧人物的详细介绍txt文本，每个txt文本以该人物的名字命名，正文是一大段详细介绍；Name\_Entities.xml是对于爬到的每一条人物介绍识别出正文部分的人名，地名和戏曲名，以xml格式存储；Opera\_characters\_Xikao\_1000.xml是从爬到的所有京剧人物介绍中提取的前1000条数据用于信息抽取；Opera\_characters\_Xikao\_All.xml是爬虫爬到的所有京剧人物数据；Relationship.txt文本是根据正则表达规则抽取出来的实体关系，主要包括师徒，人物和出生地从属关系，人物和科班/学校从属关系；Relationship.xml则是将Relationship.txt中的关系以xml格式表示；OtherMethodRelationship.txt则是利用深度学习的模型抽取出的人物关系之间的概率分布。

第二部分是Program文件夹，包含了本次作业的所有程序。WebCrawler.py是网络爬虫程序，RunNER.py是命名实体识别的程序，RelationExtraction.py是实体关系抽取的程序，ProcessMes.py是简单处理xml文本的程序；而OtherMethodForExtraction文件夹下包含了深度学习方法抽取实体关系的模型，数据和代码。

**2.实验介绍**

**2.1网络爬虫**

代码文件是WebCrawler.py文件。

通过分析作业要求中提交中的5个网址，发现中国京剧戏考（www.xikao.com）网提供的京剧人物最为充分，包含了2665名京剧人物。进入其网页观察发现，首先需要从4个子分类（生旦净丑）的网页下爬取所有人物的详细介绍网址，再

遍历每个网址去爬虫其中的人物详细信息。

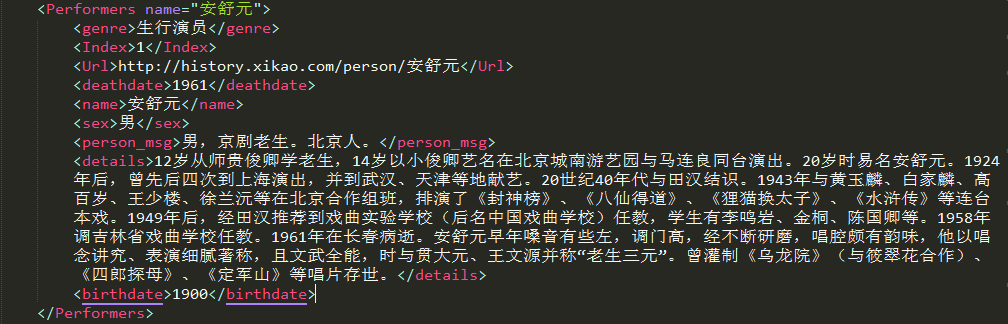
在子分类的网页下，使用如下正则表达式：

"<li class=\"bullet\">.\*?<a href=\"(.\*?)\" class.\*?>.\*?(<span class=\"brief\_info\">(.\*?)</span>.\*?)?</li>"

爬取人物详细介绍网站，出生逝世年月。之后进入详细介绍网站，爬取京剧人物姓名，性别以及详细信息介绍，利用如下正则表达式：

"<div class=\"namecard\">.\*?</div>.\*?</div>.\*?<b>(.\*?)</b>(.\*?)<br />(.\*?)<hr size=\"1\".\*?</td>"

爬取到的信息以xml的格式写入Opera\_characters\_Xikao\_All.xml文件，xml文件的格式如下图所示，



其中<genre>表示该京剧演员的行当（角色类型），<Index>是序号，<Url>是该京剧人物详细介绍的网站地址，<deathdate>是该人物的逝世日期（部分人物没有），<name>是人物姓名，<sex>是性别，<person\_msg>是人物的简单介绍，<details>是人物的详细介绍，<birthdate>是人物的出生日期。以上就是一个京剧人物爬取到的所有信息。获取其中前1000条进行后面的实体分析。

**2.2命名实体识别**

代码文件是RunNER.py文件。

使用Stanford NLP包进行姓名和地点的命名实体识别，而戏曲名则使用书名号“《》”进行正则匹配。

在进行命名实体识别之前，首先调用ProcessMes.py中的getDetailMes函数将xml中爬取到的1000条人物介绍以迭代器的方法返回。

命名实体识别使用Stanford的NLP工具包，首先在Stanford NLP官网（<https://stanfordnlp.github.io/CoreNLP/index.html>）下载最新的模型文件，包括：

（1）CoreNLP完整包（<http://nlp.stanford.edu/software/stanford-corenlp-full-2017-06-09.zip>）：下载后解压到工具目录；

（2）中文模型（<http://nlp.stanford.edu/software/stanford-chinese-corenlp-2017-06-09-models.jar>）：下载后复制到上述工作目录。

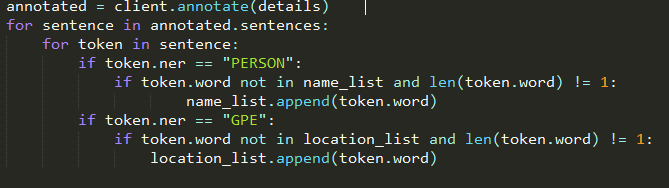
由于本项目使用python语言进行编写，而Stanford NLP是使用java语言开发，因此还需要安装stanza（<https://github.com/stanfordnlp/stanza>）的python包，用于在python环境下调用Stanford NLP工具。可以在上述github网站上下载后安装，也可以使用命令“pip install stanza”直接进行下载。

下载好上述所有工具后，在python程序中使用NLP函数之前，需要在NLP工作目录下启动cmd窗口输入如下命令启动本地服务器：

java -mx4g -cp "\*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -port 9000 -timeout 15000

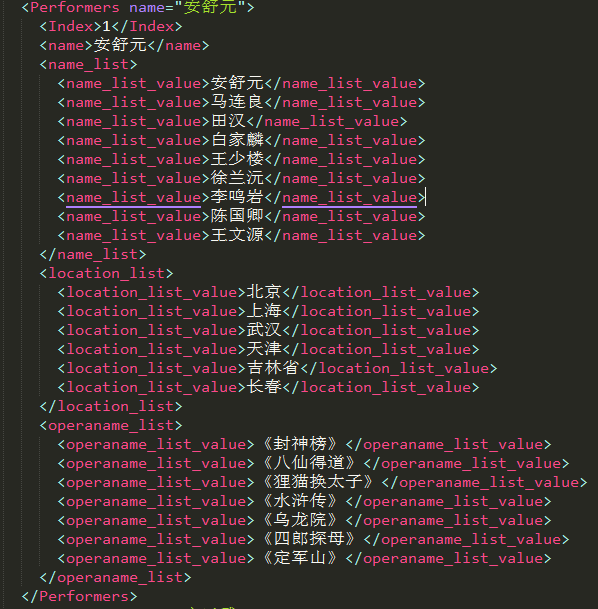
服务器启动后，在python代码中导入stanza.nlp.corenlp包中的CoreNLPClient，之后即可使用如下代码在Python中创建一个CoreNLPClinet对象，并调用该对象相应的函数进行命名实体识别。

首先进行戏曲名的识别，戏曲名识别使用书名号“《》”进行正则匹配"(《.\*?》)"，找到正文中所有戏曲名后，存入dict中，为了防止之后识别姓名和地名出现戏曲名中的姓名或者地方，存储戏曲名后使用“#”将其替代。姓名和地名识别代码如下：



循环每一个句子，判断某个句子中的词的Ner标注是否属于“PERSON”或者“GPE”，其中需要把词长度等于1的情况去除（与实际不相符）。这样就将人物正文介绍中出现的姓名和地点进行了筛选。

识别出姓名，地名以及戏曲名后，将其以xml格式存入，存入实例如下：



**2.3实体关系抽取**

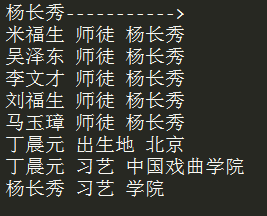
代码文件为RelationExtraction.py文件。**代码测试时输入需要进行测试的文件夹路径以及输出结果的保存路径和文件名。输入空格则表示默认输入路径为“../Data/txt\_details”，输出结果文件为“../Data/Relationship.txt”**。

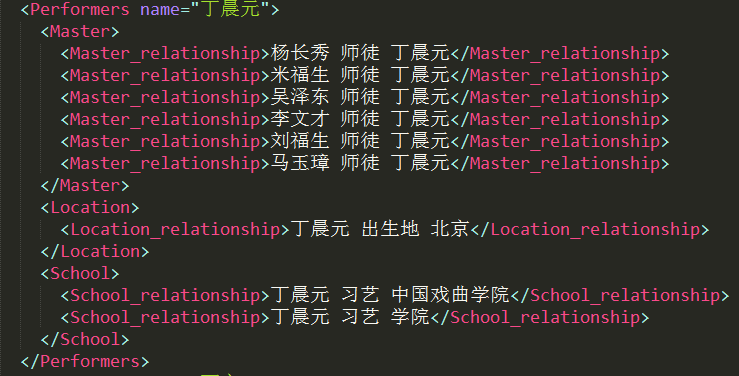
实体关系抽取使用了正则表达规则匹配的方法进行实体关系的抽取，通过观察师徒，出生地以及科班/学校的文本表达特性，总结出相应的规则，并将规则以正则表达式的方式去文本中匹配。

由于需要识别出句子中的实体类别，分句以及分词，故还需调用第二步介绍的NLP工具包，在NLP工作目录下的cmd中输入java -mx4g -cp "\*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -port 9000 -timeout 15000命令打开本地服务器，之后在python代码中调用相应函数即可。

循环正文中的每一个句子，判断是否匹配师徒关系正则表达，若匹配则找到句子中的“PERSON”标注。由于师徒关系的主语不同，故使用两个正则表达去匹配师徒关系，首先是抽取该名家的师父，正则表达式如下："(.\*师从)|(.\*(从|向).\*学.\*)|(.\*拜.\*)|(.\*被.\*收为.\*)|(.\*从师.\*)|(.\*开蒙(?!戏).\*)|(.\*师承.\*)|(.\*为师.\*)|(.\*受.\*指导.\*)|(.\*被.\*收为.\*)|(.\*为.\*弟子)|(.\*得.\*真传.\*)|(.\*传授.\*)|(.\*亲授.\*)|(入.\*习)"，而抽取改名家徒弟的正则表达式如下："(.\*培养.\*)|(.\*收.\*弟子.\*)"。

人物和出生地从属关系以及人物和科班/学校从属关系同师徒关系的处理，使用的正则表达关系分别如下："(.\*(?<!族)人)|(生于.\*)"、"(?<!调)(入|从|在|到|考入|毕业于|坐科)(.\*?(?<!的))(学校|科班|学院)(?!任教)"。处理完这3种关系后将关系写入Relationship.txt和Relationship.xml文件中，例子如下：



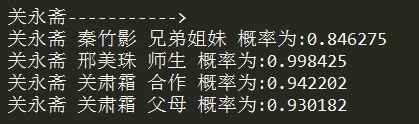


**2.4额外加分（关系抽取）**

对于第三步，在github上找到一个使用Bi-GRU和字向量做端到端的中文关系抽取的项目RE\_BGRU\_2ATT，链接如下：<https://github.com/crownpku/Information-Extraction-Chinese/tree/master/RE_BGRU_2ATT>。利用该项目训练好的模型，在1000个文本上抽取出概率大于0.8的10种关系，分别为父母，夫妻，师生，兄弟姐妹，合作，情侣，祖孙，好友，亲戚，同门以及上下级。

代码，数据以及模型位于Program文件夹下的OtherMethodForExtraction文件夹。测试时输入需要进行测试的文件夹路径以及输出结果的保存路径和文件名。输入表格表示默认输入路径为“../../Data/txt\_details”，输出结果文件为“../../Data/OtherMethodRelationship.txt”。

RE\_BGRU\_2ATT项目使用双向GRU，字与句子的双重Attention模型，输入两个姓名实体和一个句子，从句子中判断该两个实体属于以上10种关系的概率。首先遍历文本中的所有句子，若存在标注为“PERSON”的实体，则将正文的名家姓名，该实体姓名以及句子放入RE\_BGRU\_2ATT的模型中进行关系抽取，输出为概率大于0.8的10种关系之一。抽取的关系如下图所示：



**2.5心得体会**

这次作业使用爬虫爬取数据并对数据进行实体分析让我收获了很多，对如何获取数据，以及如何对数据进行分析有了更深的了解。

在短时间内学习并编写爬虫程序，Stanford NLP工具的使用以及深度学习关系抽取模型的使用都使我受益匪浅，总之，这次作业虽然花了我比较多的时间，但是最后的结果还是比较满意。