**CRF工程文档**

**开发工具的使用的方案**

Anaconda

PyCharm

Crf++工具包

**一、开发环境搭建**

**1、Anaconda**

Anaconda是一个开源的Python含了conda、Python等180多个科学包及其依赖项。Anaconda提供了包管理与环境管理的功能，可以很方便地解决多版python并存、切换以及各种第三方包安装问题，并且已经包含了Python和相关的配套工具。

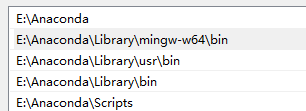
Anaconda利用工具/命令conda来进行package和environment的管理。

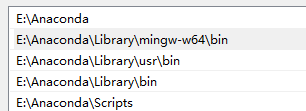
Anaconda官网提供了3.x和2.7两个版本，此工程选用的是3.x的版本，工程压缩包里有。

**安装过程**：

1、exe文件安装，可以自己选定安装位置。

2、Anaconda安装好后，我们把bin目录加入PATH（如果有就不用加了），我的环境配置结果显示如下：





3、cmd打开命令行界面，输入conda--version检查配置是否正确。如下图可以显示conda版本则说明安装成功。

****

**安装包：**

安装Anaconda之后，我们就可以很方便的管理安装包（安装，卸载）。安装包选用conda或者pip命令在dos窗口完成。此项目一般选用的pip命令。具体实现命令如下：

安装：pip install xxx(某包名字)

卸载**：**pip uninstall xxx(某包名字)

需要安装的包：结巴分词工具（jieba）

**2、PyCharm**

PyCharm是一种Python IDE，带有一整套可以帮助用户在使用Python语言开发时提高其效率的工具，比如调试、语法高亮、Project管理、代码跳转、智能提示、自动完成、单元测试、版本控制。

PyCharm官网包含专业版和社区版。专业版（Professional）需要激活，并且社区版（Community）已经包含了我们所需要的基本功能，所以这里我们选择社区版（Community），工程压缩包里也有。

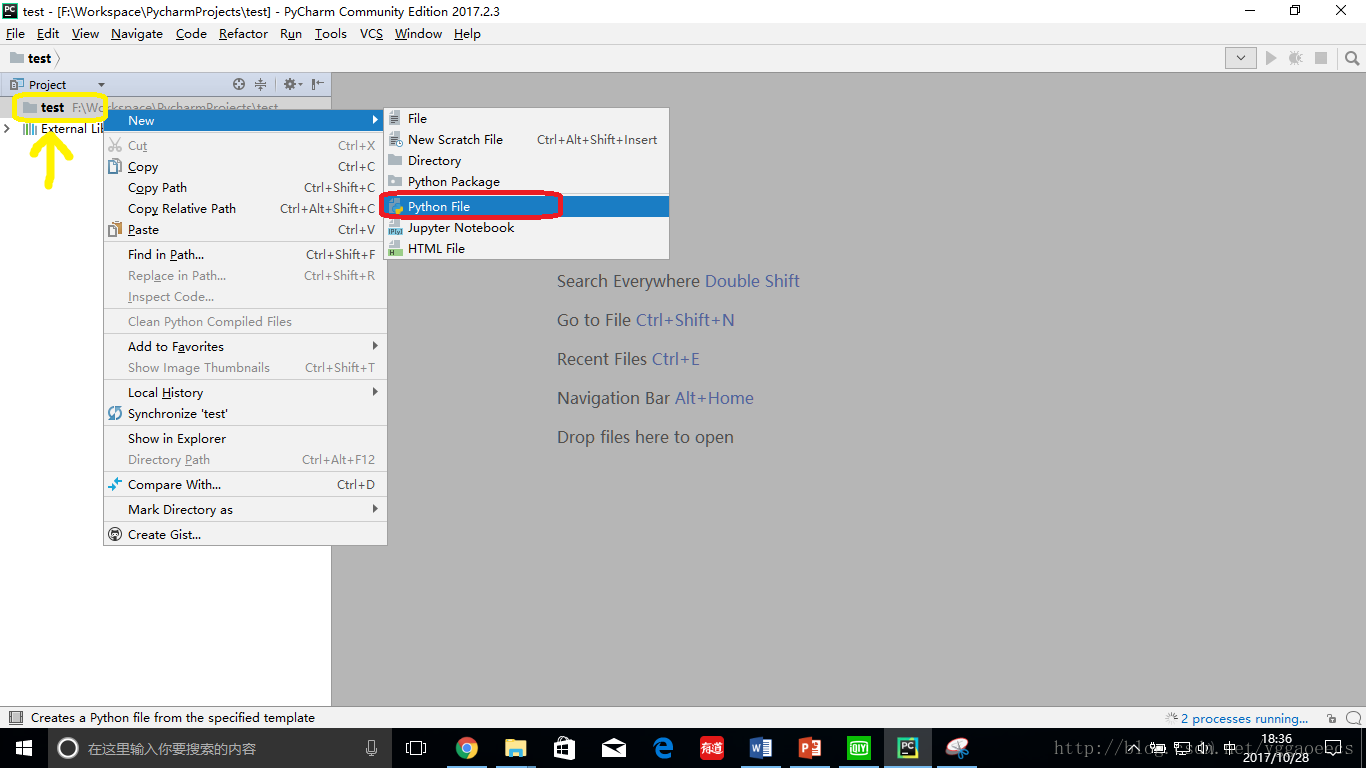
**安装过程：**

双击.exe文件进入安装界面，自己选择安装路径。按照自己电脑操作系统的位数（64位 or 32位）选择Create Desktop Shortcut，依次点击Next –> Install –> Finish 完成软件的安装。

**PyCharm的使用**

1、点击桌面上的PyCharm图标，打开如下图6所示的界面，点击OK继续。2、点击Create New Project，进入如下图8的界面。图中的Location是选择你创建python工程的位置及工程名字（根据自己的情况选择，默认位C盘），我的工程目录为C:\Users\gd006\PycharmProjects，工程名字为test（可以随便取）；图中的Interpreter是你安装Python的解释器，默认的情况下已经帮你选择好，目录为Anaconda的安装目录下的python.exe文件。选择好后，点击create。

3、进入的界面如下图所示，鼠标右键点击图中的工程名字（箭头所指处），然后new选择Python File, 在弹出的界面中填写新建Python文件的名字（任意填写）。



4、文件创建成功后，在新建的工程下面出现了你新建的python文件。之后便可以编写自己的程序了。

**3、Crf++工具包**

下载crf++0.58.zip版本，这是已经编译好的，可以直接拿来使用

说明：

解压缩crf++0.58.zip后可看到如下文件：

doc文件夹：就是官方主页的内容。

example文件夹：有四个任务的训练数据、测试数据和模板文件。

sdk文件夹：CRF++的头文件和静态链接库。

crf\_learn.exe：CRF++的训练程序。

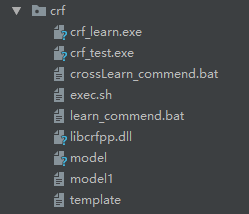
crf\_test.exe：CRF++的预测程序

libcrfpp.dll：训练程序和预测程序需要使用的静态链接库。

实际上，需要使用的就是crf\_learn.exe，crf\_test.exe和libcrfpp.dll，这三个文件。

**在上面pycharm中创建的项目（自己命名的）中创建crf文件夹**

把crf\_learn.exe，crf\_test.exe和libcrfpp.dll放入文件中。具体如下图所示（工程项目中已经完成）：



**二．Crf实验过程：**

1、准备数据。中医妇科方剂主治信息文本分为2个部分，70%的部分作为训练语料，共计5000条训练数据；30%的部分作为测试语料，共计2000条训练数据。

2、抽取标注词典。人工标注工具得到包含病名、症状、治法。证型词典的.ent文件，通过程序抽取里面的词典并分别存储到各个文件夹中。

3、训练语料和测试语料的准备。先用结巴分词器对训练语料和测试语料做分词处理，然后用我们抽取的词典对分好词的训练语料做标注，生成语料特征（crf训练的输入）。

4、然后开始进行crf训练，训练结束生成特征模板存储在model文件中。

5、用训练好的特征模板对分好词的测试语料做测试。

6、测试结果：共检测到新词 1026 个。其中测试集中已标注的新词： 658 个

准确率为 658 / 1026 = 0.6413255360623782