# 1. Stanford Word Segmenter

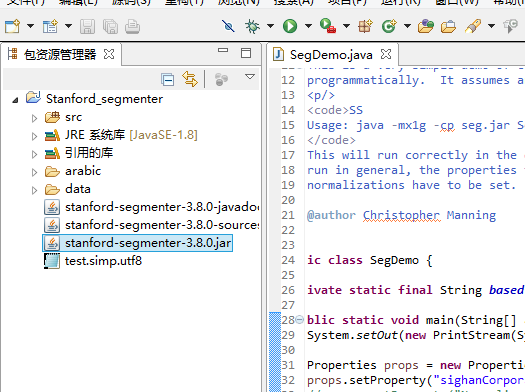
# Stanford Word Segmenter是斯坦福大学NLP group研发的一套基于CRF的开源中文分词系统，采用CRF(Conditional Random Fields)算法。下面将下载Stangford Word Segmenter 软件包,在Eclipse上调试成功，并进一步分析源码。

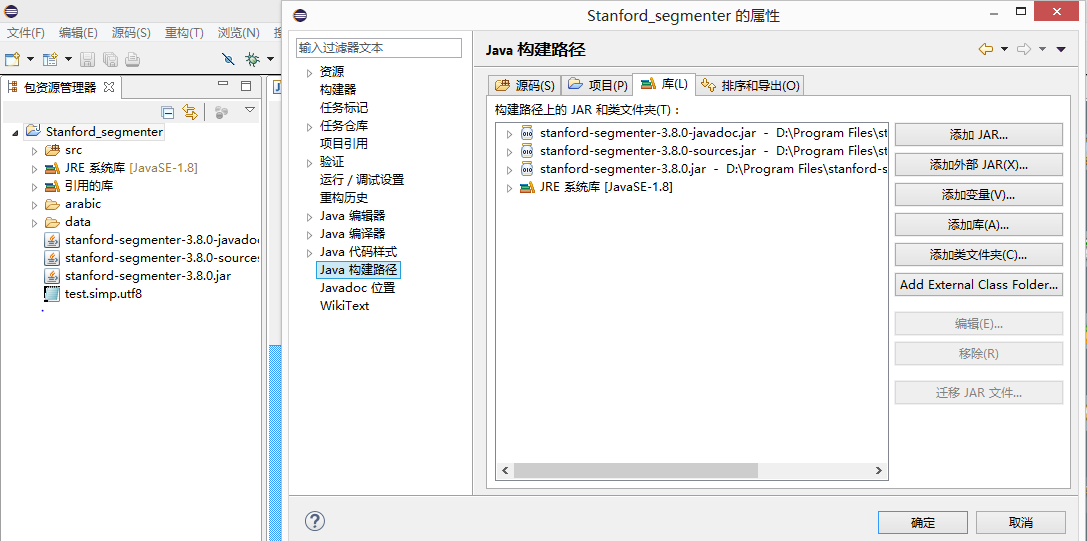
1,下载 Stanford Word Segmenter软件包；

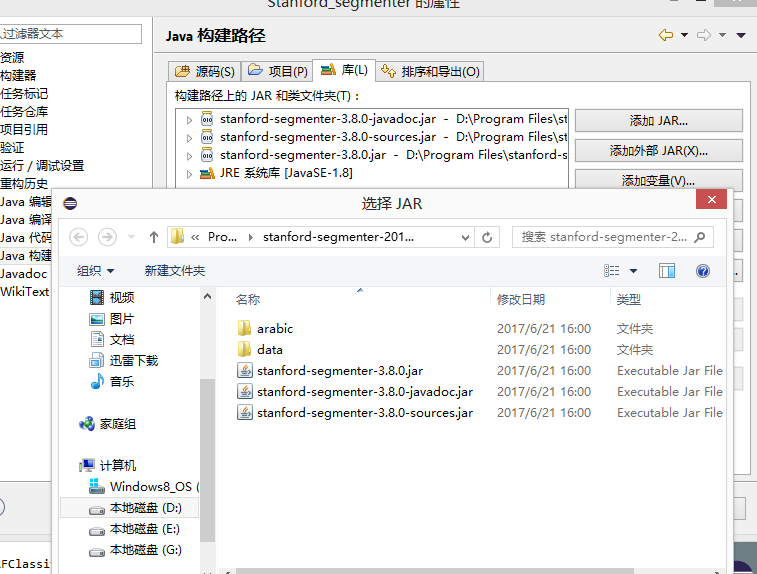
2，在eclipse上建立一个Project  Stanford\_segmenter。解压Stanford Word Segmenter软件包，将其中的data,arabic，test.simp.utf8文件夹复制到项目下。

3，添加需要的jar包，stanford-segmenter-3.8.0.jar , stanford-segmenter-3.8.0-javadoc.jar , stanford-segmenter-3.8.0-sources.jar.

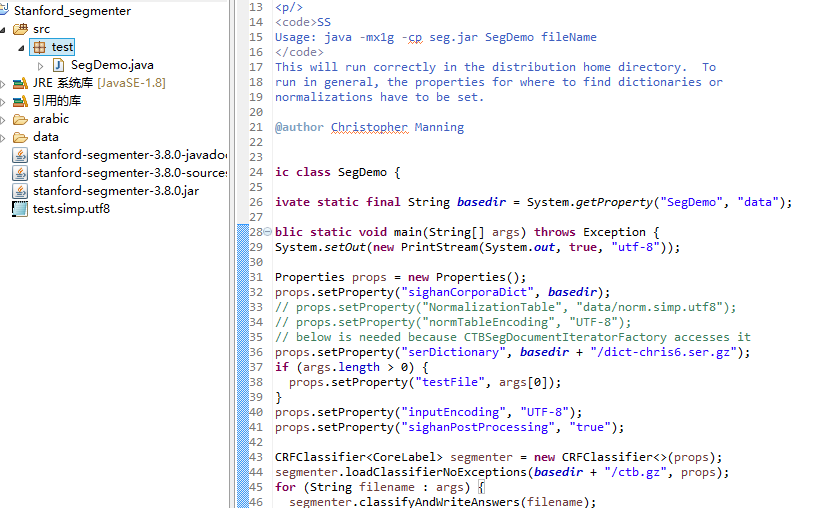
步骤：点击Project->Properties->Java Bulid Path->Libraries->Add External Jars





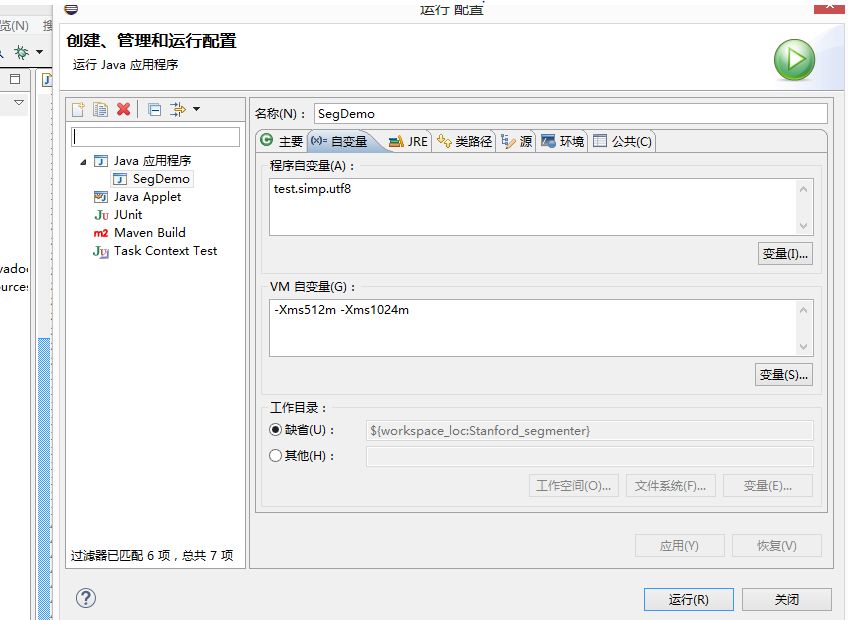


4,在项目下，建一个test包，在包下建立一个SegDemo.java，将解压出来的SegDemo的内容复制进去。



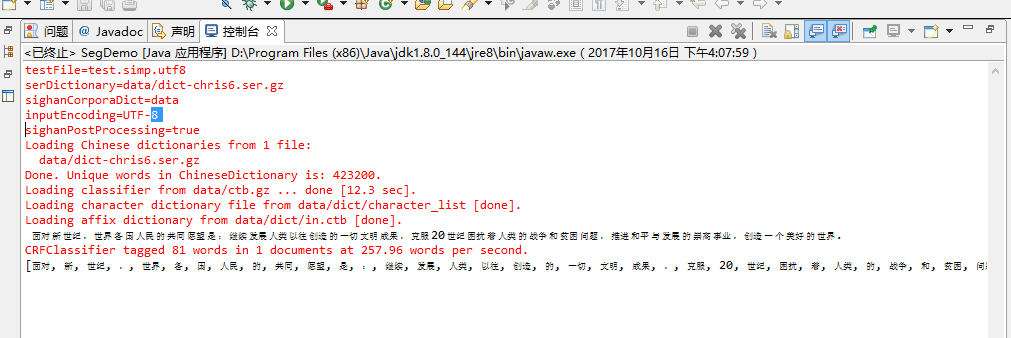
5，设置运行环境。

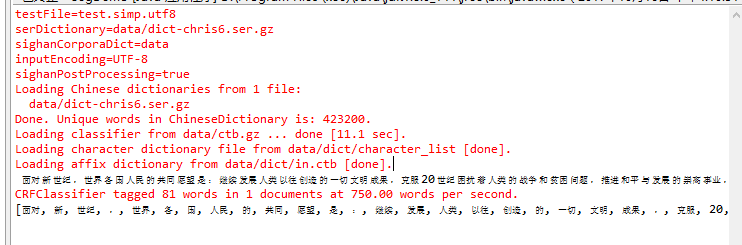
运行SegDemo，Run As-> Run Configurations,运行需要传入参数，test.simp.utf8.



由于Stanford-Sementer占用的内存比较大，所以需要设置VM arguments,不然就会超内存。

6，运行结果如下，可以看出分词的效果。





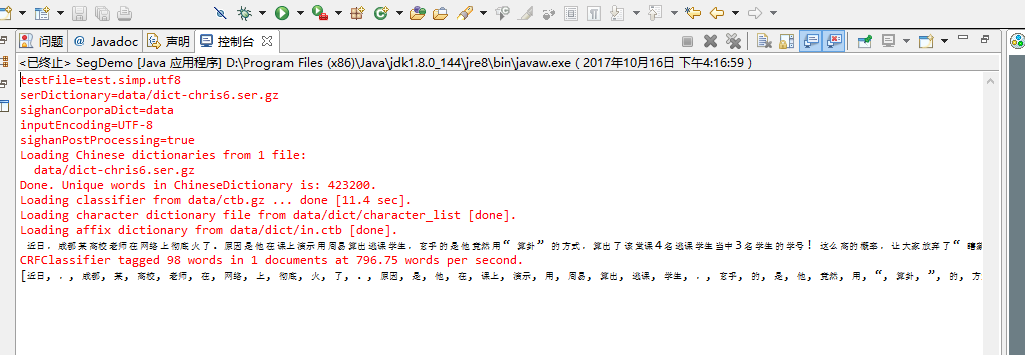
对文本进行分词结果：

面对 新 世纪 ， 世界 各 国 人民 的 共同 愿望 是 ： 继续 发展 人类 以往 创造 的 一切 文明 成果 ， 克服 20 世纪 困扰 着 人类 的 战争 和 贫困 问题 ， 推进 和平 与 发展 的 崇高 事业 ， 创造 一 个 美好 的 世界 。

分词结果：

[面对, 新, 世纪, ，, 世界, 各, 国, 人民, 的, 共同, 愿望, 是, ：, 继续, 发展, 人类, 以往, 创造, 的, 一切, 文明, 成果, ，, 克服, 20, 世纪, 困扰, 着, 人类, 的, 战争, 和, 贫困, 问题, ，, 推进, 和平, 与, 发展, 的, 崇高, 事业, ，, 创造, 一, 个, 美好, 的, 世界, .]

7.多次测试结果：

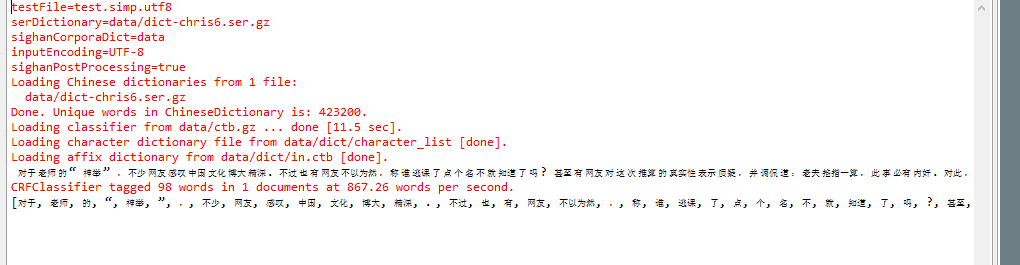


对文本进行分词结果：

近日 ， 成都 某 高校 老师 在 网络 上 彻底 火 了 。 原因 是 他 在 课上 演示 用 周易 算出 逃课 学生 ， 玄乎 的 是 他 竟然 用 “ 算卦 ” 的 方式 ， 算出 了 该 堂课 4 名 逃课 学生 当中 3 名 学生 的 学号 ! 这么 高的 概率 ， 让 大家 放弃 了 “ 瞎蒙 ” 的 质疑 。

分词结果：

[近日, ，, 成都, 某, 高校, 老师, 在, 网络, 上, 彻底, 火, 了, 。, 原因, 是, 他, 在, 课上, 演示, 用, 周易, 算出, 逃课, 学生, ，, 玄乎, 的, 是, 他, 竟然, 用, “, 算卦, ”, 的, 方式, ，, 算出, 了, 该, 堂课, 4, 名, 逃课, 学生, 当中, 3, 名, 学生, 的, 学号, !, 这么, 高的, 概率, ，, 让, 大家, 放弃, 了, “, 瞎蒙, ”, 的, 质疑, 。]



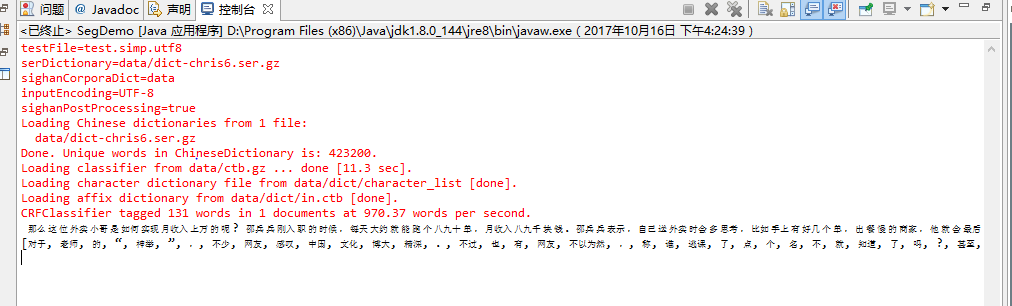
对文本进行分词结果：

对于 老师 的 “ 神举 ” ， 不少 网友 感叹 中国 文化 博大 精深 。 不过 也 有 网友 不以为然 ， 称 谁 逃课 了 点 个 名 不 就 知道 了 吗 ? 甚至 有 网友 对 这 次 推算 的 真实性 表示 质疑 ， 并 调侃 道 ： 老夫 掐指一算 ， 此 事 必有 内奸 。 对此 ， 你 怎么 看 ?

在代码中进行输入分词结果：

[对于, 老师, 的, “, 神举, ”, ，, 不少, 网友, 感叹, 中国, 文化, 博大, 精深, 。, 不过, 也, 有, 网友, 不以为然, ，, 称, 谁, 逃课, 了, 点, 个, 名, 不, 就, 知道, 了, 吗, ?, 甚至, 有, 网友, 对, 这, 次, 推算, 的, 真实性, 表示, 质疑, ，, 并, 调侃, 道, ：, 老夫, 掐指一算, ，, 此, 事, 必有, 内奸, 。, 对此, ，, 你, 怎么, 看, ?]

7、文档和代码分词不同情况



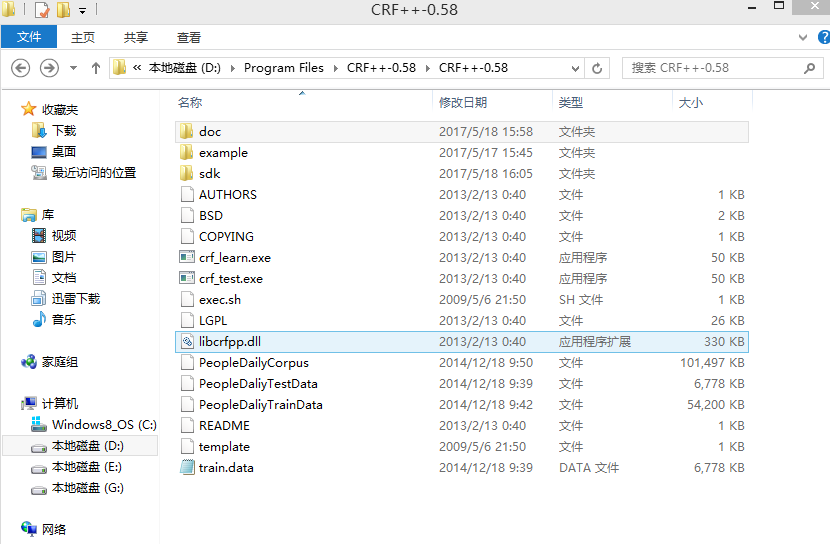
对文本进行分词结果：

那么 这 位 外卖 小哥 是 如何 实现 月收入 上万 的 呢 ? 邵兵 兵 刚入 职 的 时候 ， 每天 大约 就 能 跑 个 八九十 单 ， 月收入 八九千 块 钱 。 邵兵 兵 表示 ， 自己 送 外卖 时会 多 思考 ， 比如 手上 有 好几 个 单 ， 出 餐 慢 的 商家 ， 他 就 会 最后 一 个 去取 ， 减少 等 餐 的 时间 。 他 平时 会 注重 规划 送餐 路线 ， 送 过 的 地方 也 会 过目不忘 。

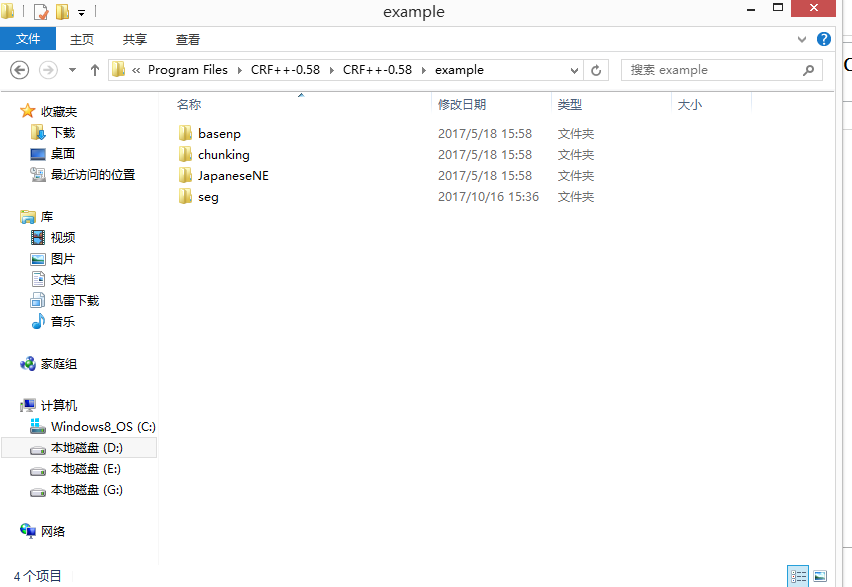
1. crf++

我下载的是CRF++0.58.zip的版本，解压。

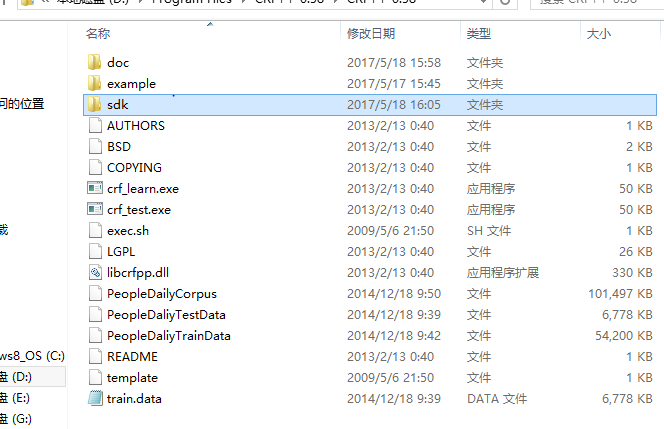
 doc文件夹：就是官方主页的内容。



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。

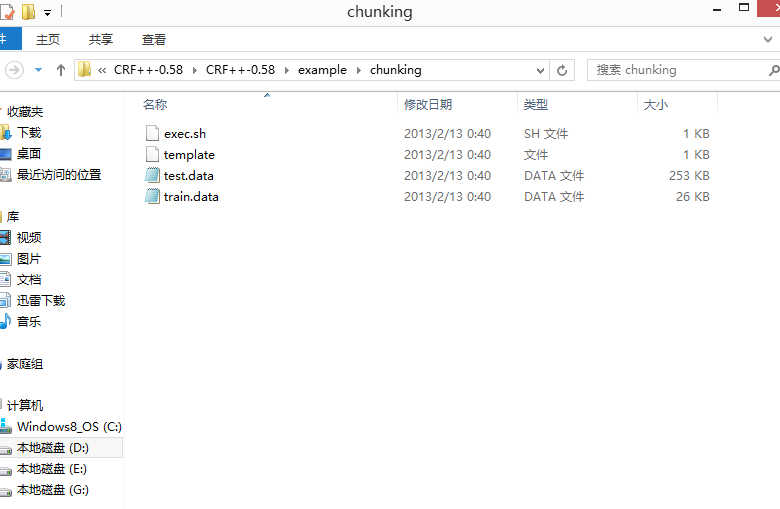


    实际上，需要使用的就是crf\_learn.exe，crf\_test.exe和libcrfpp.dll，这三个文件。

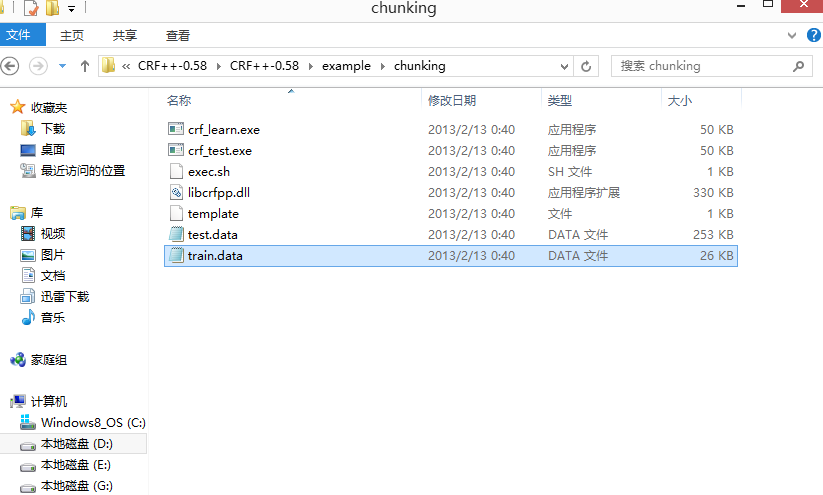
可以先拿example中的某个例子，做一下测试。例如：example中chunking文件夹，其中原有4个文件：exec.sh；template；test.data；train.data。

template为特征模版；test.data为测试数据；train.data为训练数据。

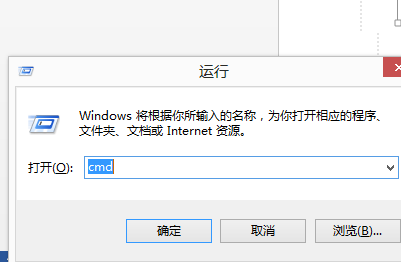
可以选择example里的某个例子做测试，比如选chunking。



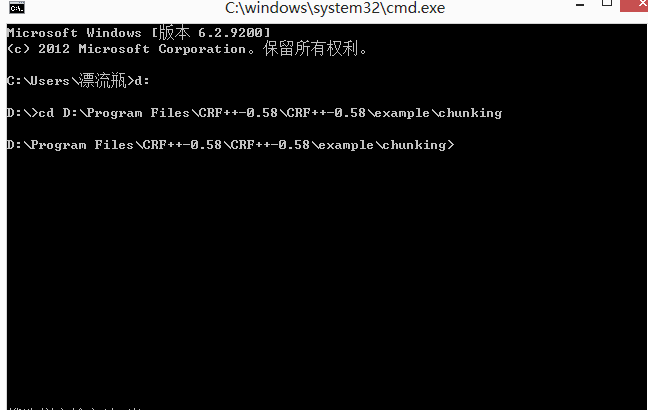
将crf\_learn.exe；crf\_test.exe；libcrfpp.dll三个文件复制到到，含有exec.sh；template；test.data；train.data的文件夹（chunking）里。

操作：

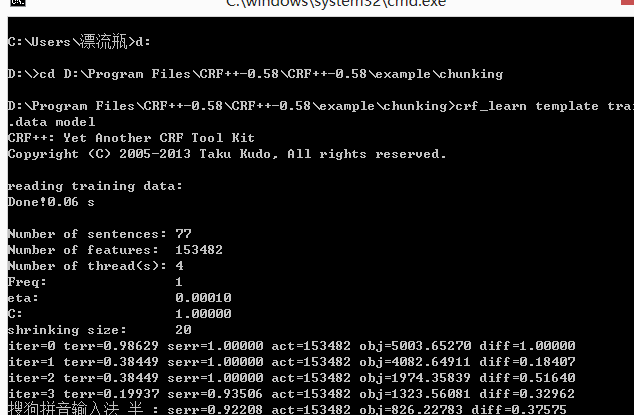
Cmd

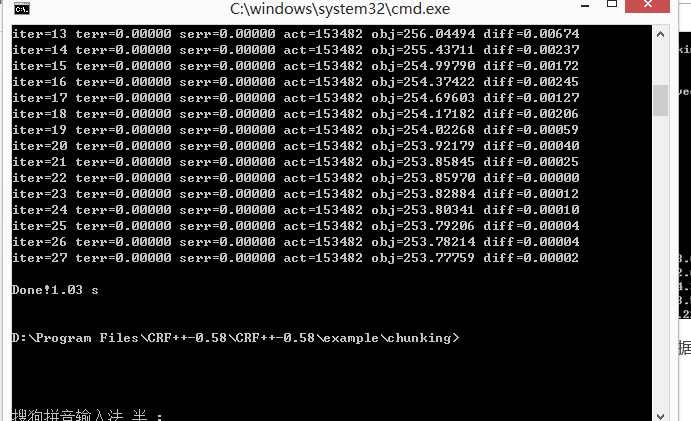


cd进入该文件夹

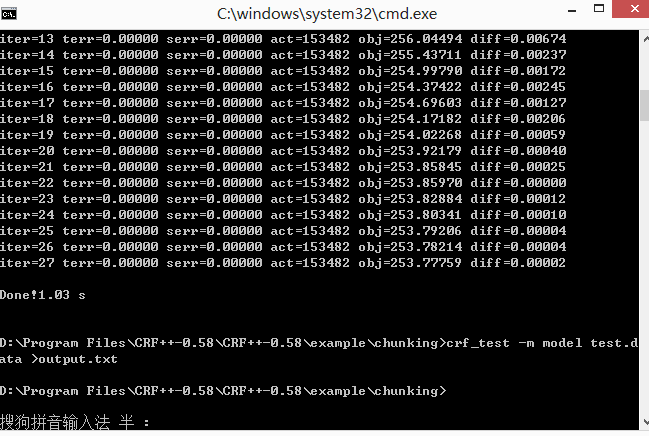


crf\_learn template train.data model   训练数据





crf\_test -m model test.data >output.txt   测试数据



生成的output.txt

