语言处理模块

知识提取：

## 当我们获得一段文本时，想要将其中的关键信息提取出来的话，首先要进行分词处理，我们采用jieba库进行分词。

第一步，安装jieba：

jieba支持pip或者conda安装，直接在命令行执行：

pip install jieba

第二步，确定使用模式：

Jieba有全模式和精确模式等模式，这里我们采用精确模式（默认即为精确模式）。

第三步，导入自定义词库：

因为我们要对歌名和人们进行分词处理，极其容易出现分词错误，所以再这里要先提前创建自定义词库，我们的自定义词库是通过爬虫将网络上的歌名、歌手、作词人、作曲人和编曲人爬取下来，最终创建自定义词库，具体详细内容可见.......

当我们将jieba的配置都设定好之后，开始分词，这时就可以看到分词的效果很好：

原始文本：《晴天》是周杰伦作词、作曲、编曲并演唱的歌曲，收录在周杰伦2003年7月31日发行的专辑《叶惠美》中。

分词之后：《 晴天 》 是 周杰伦 作词 、 作曲 、 编曲 并 演唱 的 歌曲 ， 收录 在 周杰伦 2003 年 7 月 31 日 发行 的 专辑 《 叶惠美 》 中 。

然后，要建立关键词之间的联系。我们将“作词”、“作曲”、“编词”、“编曲”和“演唱”之类的词作为关键词，当一句话之中出现这些词的时候，我们就需要寻找这一句话中出现的人名，当同时找到这些信息时，即可建立关系。同时，我们还需要将歌曲名和专辑名找出来，只要是被书名号包围的即为歌曲名与专辑名。当找到这些信息后，我们即可把这些关键信息链接起来，输出出来。效果如下：

原始文本：《晴天》是周杰伦作词、作曲、编曲并演唱的歌曲，收录在周杰伦2003年7月31日发行的专辑《叶惠美》中。

处理之后：

《晴天》：

周杰伦演唱

周杰伦作词

周杰伦作曲

周杰伦编曲

收录于《叶惠美》

将这些操作重复进行即可实现多组文本的提取。

知识合并：

当我们获得到一些处理后的信息后，可以将这些信息中重叠的地方合并起来，例如，周杰伦演唱了《晴天》，周杰伦演唱了《青花瓷》，我们可以把这些信息和起来成：周杰伦演唱了：《晴天》，《青花瓷》。这样就可以将所有的信息合起来。