Yoshikoder PC版上手指南

王明德

Yoshikoder 出自Harvard Weatherhead Center的Identity Project，由诺丁汉大学的Will Lowe主持开发，是一款跨平台、多语种的内容分析工具。它适用于ASCII、Unicode、或者国标码下的文本文件（TXT），允许使用者自主编撰关键词词库，“装载”文档；进行分类词频统计或者“词汇-语境”对照等基本的内容分析。本指南将简要介绍如何在PC操作环境下安装和使用Yoshikoder，及中文报纸内容分析的基本流程。

一． 下载和安装

首先，请进入此页面: http://sourceforge.net/project/showfiles.php?group_id=167731 ，下载Yoshikoder (0.6.3-Preview.3.)。 其下方另一个小工具Yoshikoder-converter (0.2.1.1)的功能是将非TXT文档（PDF、DOC、HTML）转换为UTF-8格式的文本文档，虽然目前尚无用处，但是也可以保存下来，以备不时之需。 其次，为了使Yoshikoder能够识别中文，还需要装载一个叫 ”SCTokenizer.jar” 的 Java 插 件 。 该 文 件 可 以 在 这 里 找 到 ： http://yoshikoder.sourceforge.net/resources.html 。下载Tokenizer之后，运行Yoshikoder，在<File>菜单下选择<Preferences>。 然后在弹出框中点选<Tokenizer>标签，点击 ”Add” 将刚才下载的SCtokenizer.jar文件加入，安装即告完成。

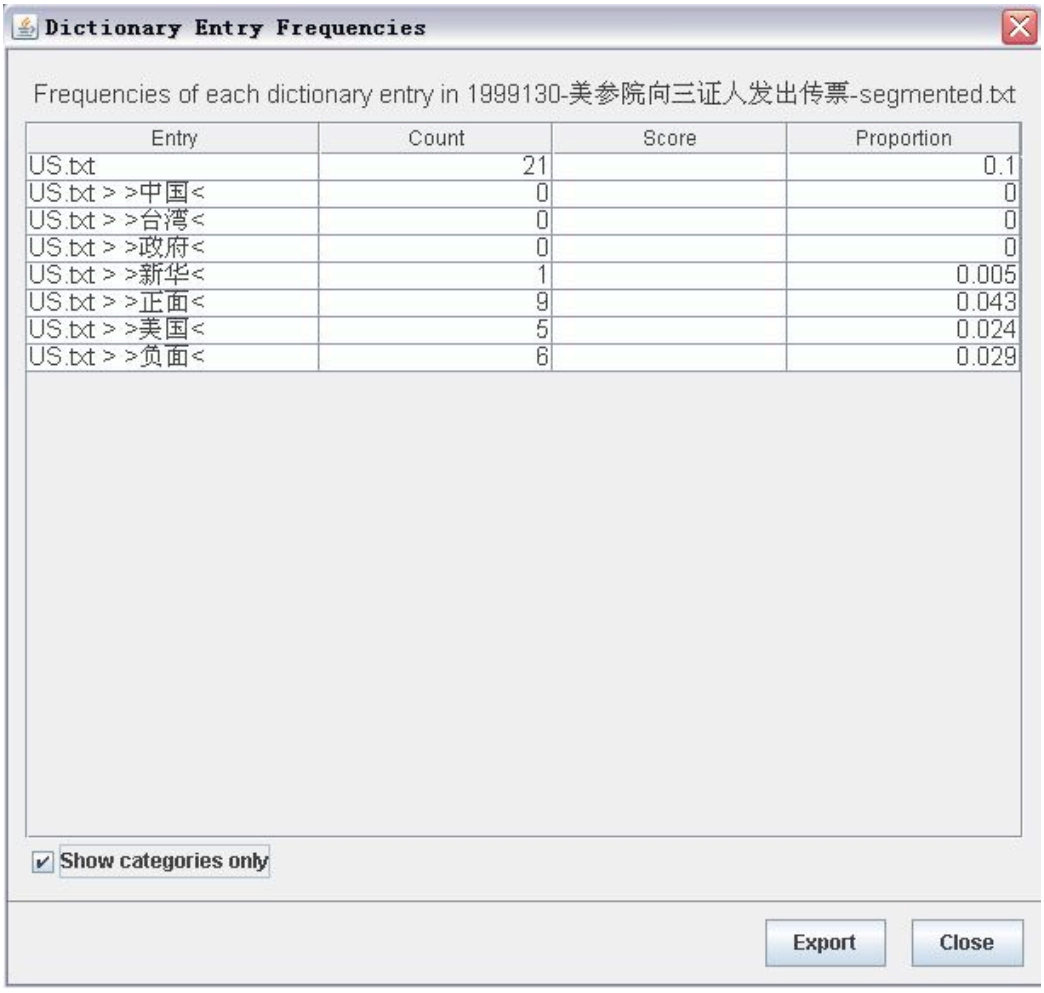
二． 基本操作

Yoshikoder的基本功能是将目标文本中的关键词与事先编制好的“关键词分类词库”，进行比对，统计各分类项下关键词“相符”的频率。操作界面如下：

分析一篇文章之前，必须先载入或者手动编制“关键词分类词库”(Dictionary File，YKP文件)，然后载入要分析的TXT文本文档。注意，相反的操作顺序会导致当前载入的全部文档消失。由于文件载入速度可能相当慢(取决于不同机器)，发生这种情况有可能浪费很多时间。词库由“分类”（category）和下属的 “词条”（pattern）组成，可以用右键弹出菜单添加/编辑/删除。然后就要将文本本档载入Yoshikoder。实现的方法有两种。 目前请用“Import Document”的方式。具体操作：点击<Document>菜单，选择<Import Document>，选中TXT文件后在弹出框里将"Encoding"设置为UTF-8；相应的文件就会出现在正中的浏览器内。其右边的列表栏则会显示当前载入的所有文本文件的名称。

Yoshikoder最重要的两个功能，“高亮标识关键词”（Hignlight）以及“关键词语镜排布”(Concordance)，都和词库的设置直接相关。左键选择词库栏中的某一分类或者词条，再点击上方工具栏里的“放大镜”图标，浏览框内的文本中如有与所选分类或词条相同的关键词，就会被标示为黄色高亮状态（见图示：“新华社”）。选择关键词后点击“语境”排布（‘三层抽屉’图标），最下方的浏览框内将显示全部被选关键词在目标文本中的语境。范围为该关键词前后各八个单词以内。

接下来要用到的功能是“词频统计”（Report）中的<Dictionary Report>和<Concordance Report>。前者是按照关键词词库的分类，统计目标文本中所有相符关键词的数量和比例；后者的功能相同，目的则是为了统计被选关键词所在的语境中，与词库相符的关键词频率。两者的具体操作和界面都相同：点选<Report>下相应的选项，待弹出框现身后，勾选其下方的“Show Categories Only”。如图所示，Yoshikoder将依次显示与词库相符的各类关键词的数量和比例。



(例：全部关键词US.txt = 21， Proportion = 0.1；分类关键词 >新华< = 1, Proportion = 0.005)

三. 内容分析流程

介绍了基本操作，接下对内容分析的流程稍作提示。本研究的目的是对中国报纸如何塑造“美国形象”进行分析。基本流程即阅读目标文本，统计相关的词频和数据，然后将结果记录、编码成各项Variable，然后写入数据库中。需要被编码的每一项Variable，都在Excel格式的Dataset里附有注释。请直接参考即可。

这里需要特别提示的是，在阅读的过程中要注意是否有mismatch的现象，即应该标出而没有标出的关键词；或者被错误标注以及分隔的关键词。另一个需要注意的问题则是“错误归类”。这是由语言环境的丰富变化造成的。例如“中国”分类下的“全国”词条，在很多描写他国的报导中亦会被归入“中国”一类里，类似问题只能靠人工阅读来排除。