使用说明

Java Version:

1、TextExtract.java 网页正文抽取类（不需要任何第三方包，如 DOM4j, HTML Parser 等, 代码去注释后只有百行）；
2、UseDemo.java 调用示例，注释里有详细的使用说明。

Perl Version:

****
prerequisite（实现时完全可以不用这些第三方包，只是为了方便和展示）
Win32 下：
1、安装 ActiveState Perl(Version 5.8.0 或以上)
2、打开 cmd，在 ppm 下安装以下 packages，具体命令（需要联网）：
ppm install HTML::Parser
ppm install XML::DOM
ppm install XML::Twig
Linux 下：
1、在 CPAN 下安装以下 packages，具体命令（需要联网）：
cpan
install HTML::Parer
install XML::DOM
install XML::Twig
****

本代码实现了下面两类功能：
1、测试一个网页是否可以被抽取正文
（我将网页分为两类：主题类和目录类。主题类网页像新闻、博客等，可以抽取正文；目录类像新浪首页等，不予抽取）
2、将抽取到的正文保存下来，可以直接将其保存为文本文件或 XML 文件

将 HTML 抽取正文并保存为 XML 文件，需在 cmd 下运行下面命令：
TextExtract_from_URL_to_XML.pl url.data(本文件是用来测试正文抽取的 URL 地址)

将 HTML 抽取正文并保存为.txt 文件，需在 cmd 下运行下面命令：
TextExtract_from_URL_to_TXT.pl url.data(本文件是用来测试正文抽取的 URL 地址)