项目一：搜索引擎文本预处理

要求：

- 通过下载引擎（Web Crawler/Spider）自动下载至少**500个英文**文档/网页，以及**500个中文**文档/网页，越多越好，并保留原始的文档/网页备份（如：News_1_Org.txt）
- 编程对所下载文档进行自动预处理：
  - 将各个单词进行字符化，完成删除特殊字符、大小写转换等操作
  - 调研并选择合适的中文分词技术和工具实现中文分词
  - 删除英文停用词（Stop Word）
  - 删除中文停用词
  - 调用或者编程实现英文Porter Stemming功能
  - 将中文文档进行字符化，即可被搜索引擎索引的字符单元
  - 对于**英文**文档，经过以上处理之后，将经过处理之后所形成简化文档保存（如：News_1_E.txt），以备以后的索引处理
  - 对于**中文**文档，经过以上处理之后，将经过处理之后所形成简化文档保存（如：News_1_C.txt），以备以后的索引处理

- **项目成绩：本项目成绩满分15分，直接计入总成绩**
- **项目完成日期：2020年5月11日星期一下午16：00之前交到X9422。**
- **内容要求：**详细描述实现过程、步骤、以及适当的屏幕截图，中间结果形式。
- **提交形式：**

  电子版：如项目截止时，因疫情尚不能按时返校，按电子文档pdf形式提交。命名：学号_姓名.pdf，提交邮箱后续另行通知。

  纸质版：如项目截止时，已正常返校，纸质版提交。封面填上姓名和学号。

注意事项：

- **严禁抄袭剽窃其他人成果或程序，一经查实，将视为零分。**
- 可使用任何一种你熟悉的编程语言实现：C、Java、C#、Perl、Python等。
- 尽可能将所有步骤集成为一个完整的自动化系统。如暂无法将各步骤集成为一个完整系统，则需编程实现各个步骤（模块），然后利用中间文件传递结果。
- 可利用现用有下载引擎，也可自己编写程序实现自己的下载引擎。编程实现该功能，成绩上适当加分。
- 尽量下载处理更多的网页，以备后期搜索使用。
- 可利用提供的停用词表，或者自己根据不同应用所生成的停用词表。