20181121 作业

1、 http://news.sina.com.cn/hotnews/20081031.shtml 是“新浪每日热点新闻”，其中 2008 是年份，10 是月份，31 是日期(如果月份为单数比如：7 月，则写成 07，日期同理)。请采集该链接下 20081001-20081031 之间的“国内新闻”、“国际新闻”、“社会新闻”、“体育新闻”、“财经新闻”、“娱乐新闻”、“科技新闻”、“军事新闻”【注：按点击量排行采集，此处采集包括链接对应的内容】;

2、 根据“新浪每日热点新闻”中的链接采集新闻内容，重复链接仅采集一次，采集时注意不要采集公共链接；即每个“新浪每日热点新闻”中都会出现的链接，比如：网页底部的新浪简介以及与之类似的链接；

3、 解读每个网页，将新闻实体内容取出，包括新闻标题、新闻来源、发布日期、正文等；【以下题目以第 2 和第 3 要求为前提】

4、 统计每日新闻网页中含有图片网页数量，形成每日新闻图片数量统计图，并用 Excel 制作成图；

5、 根据热点新闻目录，统计各个媒体新闻量，并用 Excel 制作成图；

6、 统计每日新闻的字数，并用 Excel 制作成图；

7、 删除所有新闻网页(新闻内容网页)中的所有 html 元素信息删除(可以简单理解为限制在尖括号之间的内容为 html 元素信息)，并将所有新闻文本集中在一个文本文件（news20081001-20081031.txt）之中；

8、 提交内容：程序代码文件命名为 20181121_学号.py、Excel 文件(命名为 20181121_学号.xlsx，图片和数据包含其中，含每日新闻图片数量、每日新闻媒体数量发布数量、每日新闻字数);

9、 提交截止日期：20181205 日 AM06:00。