哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

信息检索

实验一：网页文本的预处理

School of Computer Science and Technology

Harbin Institute of Technology

1 实验目标

本次实验目的是对信息检索中网页文本预处理的流程和涉及的技术有一个全面的了解，包括抓去网页、网页正文提取、分词处理、停用词处理等环节。本次实验所要用到的知识如下：

- 基本编程能力（文件处理、网页爬取等）
- 分词、停用词处理

2 实验环境

编程语言为：推荐使用 Python，爬虫工具推荐 urllib2，HTML 文件处理推荐 beautiful soup
其他无特殊要求

3 实验内容及要求

3.1 网页的抓取和正文提取

任务描述：通过爬虫工具爬取网页（至少 1000 个，其中包含附件的网页不少于 100 个，**多线程实现爬虫可加分**），然后提取网页标题和网页正文，以及网页中的附件并保存附件到本地，然后将附件名称记录在 file_name 字段中，**附件必须是文本文档（txt、doc、docx、xlsx 等）而不能是图片**。网页正文和网页标题可以自行定义，但一般应该是网页中你最关注的内容。例如在一般的新闻网页上，就以新闻标题为网页标题，新闻内容为网页正文，而其他诸如导航栏、广告等都是不关心的内容。为了保证可读性，网页正文中不应该包含太多 HTML 标签（如<p>、<img>等），同学们可以通过任何方法来去除掉这些标签。将爬取下来的数据保存为 json 格式，具体格式如下：

```
{
    “url”：“http://today.hit.edu.cn/article/2019/03/25/65084”，
    “title”：“计算机学院召开第 3 次科创俱乐部主席联席会”，
    “parapraghs”: "text paragraph"
    “file_name”: [file_1, file_2, … file_n]
}
```

注：请使用专门处理 json 文件的库进行处理，保证**每行**是一个标准的 json 格式数据。

提交要求：提交程序（craw.py 或 craw 文件夹）。

3.2 分词处理、去停用词处理

任务描述：将提取的网页文本进行分词和去停用词处理，并将结果保存。分词工具推荐使用由我校社会计算与信息检索研究中心开发的语言技术平台-LTP， LTP 的 Python 封装为 pyltp，这里是参考文档。停用词表采用由我校社会计算与信息检索研究中心发布的停用词表(stop_words.txt)。最后将经过分词和去停用词后的结果保存，格式如下：

{

"url"： "http://today.hit.edu.cn/article/2019/03/25/65084"，

"segmented_title"：["计算机学院"， "召开"， "第 3 次"， "科创俱乐部"， "主席"， "联席会"],

"segmented_parapraghs" :[segmented text paragraph 1],

"file_name" : [file_1, file_2, … file_n]

}

注：如果该页面存在文档，则将文档下载并将文档名称保存在 file_name 字段中。这些数据及文档同学们一定要妥善保存，后面实验 3 中将会用到。

提交要求：提交完整程序(segment.py 或 segment 文件夹)和处理后文件的**前 10 行**(preprocessed.json)。

4. 实验提交

本次实验的实验报告请严格按照"信息检索实验报告模板.docx"的格式完成，并**导出为 PDF 格式**，按"学号_姓名_实验 1 网页文本预处理实验报告.pdf"命名（例如 1140310421_张三_实验 1 网页文本预处理实验报告.pdf）。

请同学们按照各个实验模块的提交要求**正确命名提交文件，**然后将所有文件（实验报告、程序、处理后的文件）打包命名为"学号_姓名_实验 1 网页文本预处理.zip"（例如 1140310421_张三_实验 1 网页文本预处理.zip），发送到邮箱 :。

最后提醒同学们，报告或代码发现抄袭现象，该实验部分将按 0 分处理。