

第八次课外作业

正则表达式编程：

- 1、 如下图为我校校园网上新闻网页中的一区域，对应的部分 html 网页代码在 news.txt 文本文件中，请用记事本打开此文件，将所有内容作为一个字符串数据，在 Python 编程环境中 赋予一个字符串变量，再使用正则表达式提取出 14 条新闻的标题、该标题的超链接以及后方的发布日期。

输出格式为：



| | |
|----------------------------------|--------------|
| 医大要闻 | 当前位置：首页>医大要闻 |
| 我校举行红十字会遗体（器官）捐献中心安医大接受站建站二十周... | 2023-04-06 |
| 学校多个团支部和社团获省级荣誉称号 | 2023-04-06 |
| 我校斩获合肥市首届大学生反诈主题辩论赛冠军 | 2023-04-06 |
| 总人数创新高 我校8位学者入选2022“中国高被引学者”榜单 | 2023-04-06 |
| 我校举办第八届生理学知识竞赛 | 2023-04-04 |
| 第二临床医学院举办2023届毕业生春季校园双选会 | 2023-04-04 |
| 药学院专场招聘会助力毕业生高质量就业 | 2023-04-04 |
| 黄新美到附属口腔医院专题调研纪检工作 | 2023-04-04 |
| 学校召开中层领导班子换届工作动员大会 | 2023-04-03 |
| 校纪委围绕纪检监察干部队伍教育整顿督导检查工作进行部署培训 | 2023-04-03 |
| 学校召开食品安全工作会议 | 2023-04-03 |
| 【学雷锋在行动】“星翼筑梦”志愿服务队为星星点灯，将温... | 2023-04-01 |
| 我校第十届“我的中国梦”人人演讲大赛学子奋斗之志 | 2023-03-30 |
| 我校获评全国医学英语水平考试优秀考点 | 2023-03-30 |

每页 14 记录 总共 9434 记录 第一页 << 上一页 下一页 >> 尾页 页码 1/674 ☐ 跳转到

新闻标题 1: XXXXXXXXXXXX

标题超链接: XXXXXXXXX

发布时间: XXXXXXXX

新闻标题 2: XXXXXXXXXXXX

标题超链接: XXXXXXXXX

发布时间: XXXXXXXX

.....

- 2、 如右图所示为携程网上一扬州美食网页，对应的 html 网页数据存放在”扬州美食.txt”文件中，用记事本打开此文件，将所有内容作为一个字符串数据，在 python 编程环境中赋予一个字符串变量，再使用正则表达式提取出六个“美食名称”、每种美食的介绍，及去哪里吃的“店铺名称”。

输出格式为：



必吃美食

| | | |
|---|---|--|
|  <p>扬州灌汤包 扬州灌汤包制作精细，深受当地人的青睐。它以糯粉制成，用骨髓汤打肋条肉为馅，...</p> <p>哪里吃： 富春(沃尔玛店)、卢氏古宅、冶春(珍...</p> |  <p>大煮干丝 赫赫有名的大煮干丝又称喝干丝，是一道清爽健康的佳肴。这道菜对刀工的要求...</p> <p>哪里吃： 富春茶社、锦春大饭店(东关街店)、富...</p> |  <p>扬州炒饭 扬州炒饭又称扬州蛋炒饭，在世界上都是赫赫有名的佳肴。其品种丰富，有金裹银、...</p> <p>哪里吃： 恒园饭店 餐厅、富春(沃尔玛店)、食...</p> |
|  <p>狮子头 狮子头在淮扬菜系中赫赫有名，口感松软微甜，肥而不腻。这道菜做起来十分讲究，...</p> <p>哪里吃： 富春(沃尔玛店)、恒园饭店 餐厅、淮...</p> |  <p>千层油糕 千层油糕是一道历史悠久的扬州名吃，与翡翠烧麦并称淮点中的“扬州双绝”。其口...</p> <p>哪里吃： 富春茶社、晋家桥饭店(广陵路店)、...</p> |  <p>三丁包子 三丁包子是扬州的名点之一，采用鸡丁、肉丁、笋丁制成。包子皮吸收了馅料的汤汁...</p> <p>哪里吃： 富春茶社、富春(沃尔玛店)、冶春(珍...</p> |

美食名称: XXXXXX
美食介绍: XXXXXXXXXXXXXXXX...
店铺名称: XXXXXXXX

美食名称: XXXXXX
美食介绍: XXXXXXXXXXXXXXXX...
店铺名称: XXXXXXXX

.....

每一种美食都有多个店铺可以提供, 尝试提取出每一种美食可提供的所有店铺, 重新输出, 输出格式为:

美食名称: XXXXXX
美食介绍: XXXXXXXXXXXXXXXX...
店铺名称: XXXXXXXX XXXXXX XXXXXX....

美食名称: XXXXXX
美食介绍: XXXXXXXXXXXXXXXX...
店铺名称: XXXXXXXX XXXXXX XXXXXX....

.....

- 3、用记事本打开“五哈弹幕.txt”高维数据文件, 将所有内容作为一个字符串数据, 赋予一个字符串变量。

(1) 使用正则表达式提取出每个弹幕的评论(content 后面的内容)及评论的 ID 号(commentid 后面的数字)。输出评论内容(字符)最多的前五条记录:

输出格式为:

评论 ID: XXXXXXXXXXXX, 评论内容: XXXXXXXXXXXX

评论 ID: XXXXXXXXXXXX, 评论内容: XXXXXXXXXXXX

.....

(2) 将所有评论内容合成字符串, 进行分词和词频统计分析, 并作出词云图, 以证实你的分析结果。

- 4、用记事本打开 zlgwfzzp.json 文件, 将整个内容通过复制、粘贴, 赋值给一个字符串变量 words, 再使用正则表达式, 分别提取所有招聘单位的单位名称、招聘岗位名称、薪水、工作城市名及岗位待遇信息, 并按下列格式每五行输出一个单位的招聘信息然后空一行, 再输出下一个单位的信息。

输出格式为:

单位名称: XXXXXXXXXXXX

薪水: XXXXXX

岗位名称: XXXXXXXXXXXX

工作城市: XXXXXX

岗位福利: XXXX

.....

建议大家尝试使用 open() 函数来打开文本文件, 将文件内容传递给字符串变量。