

Lecture2 Challenge

本部分为已经掌握requests爬虫的同学们设计。主要为Selenium爬虫方法的实践，难度跨度较高。

问题描述：

好大夫在线是中国著名的在线医疗问诊网站，存有大量病人和专家团队沟通的对话记录，下面是一个网页例子：[https://m.haodf.com//doctorteam/flow\\_team\\_6480090760.htm](https://m.haodf.com//doctorteam/flow_team_6480090760.htm)

07:16

尚小珂 副主任医师

问题比较多，肯定需要手术。从彩超上看最大的问题是肺动脉重度返流，导致了肺动脉主干扩张，右心室扩张（右心衰），以及次要问题室间隔缺损补片残余分流，潜在的肺动脉分支狭窄。后面几个倒是小问题可以不管，肺动脉大量返流是导致右心衰的根本以及是独立的死亡原因之一，需要尽快处理。手术方式有两种，一种是老办法开刀，一种是做微创介入，目前国内可以做微创介入的只有我们这里，并且介入瓣膜数量有限已经不多了，尽快来胡教授这里就诊。

郑重提示：线上咨询不能代替面诊，医生建议仅供参考！

07:16

尚小珂 副主任医师

尚小珂医生申请@胡志伟医生查房

郑重提示：线上咨询不能代替面诊，医生建议仅供参考！

07:16

胡志伟 主任医师

你好，TOF术后会在远期出现肺动脉瓣关闭不全，导致右心功能不全，有临床症状需要干预肺动脉瓣！

郑重提示：线上咨询不能代替面诊，医生建议仅供参考！

07:16

患者

胡主任，现在的情况如何处理？需要手术吗？

现在需要你抓取这个网页上的对话信息，内容病人/医师的对话文本、不同回答医师的名称。要求以消息记录的形式存储到本地，如：

A	B	C	D	E	F
医生姓名	医生文本	患者	患者文本		
高志冬 副主任医师	您好，首先看目前的穿刺活检病理考虑胃肠间质瘤可	患者	好大夫在线友情提示：请详细描述或拍照上传病历资料，以便医生了解病情，做出更		
高志冬 副主任医师	基因检测结果具体是什么？如目前腹胀不适，首先	患者	医生，已经做过基因检测，结果是可以口服甲磺酸伊马替尼片。恩替卡韦一直在服用。		
高志冬 副主任医师	目前看基因检测kit外显子11突变，一般使用伊马替	患者	请帮忙拿下治疗方案		
高志冬 副主任医师	病情摘要/建议：病情摘要及初步印象：腹腔肿瘤 胃肠间质瘤	患者	kit外显子11突变；	总结建议:继续目前伊马替尼400mg/d治疗，定期复查血常规	

注意，该网站并不会一次性显示所有对话内容，会有一个“加载更多”的按钮，点击后才会显示一部分新对话，需要不断点击，直到该按钮消失，页面上才会显示完整的对话内容：

07:16

胡志伟 主任医师

介入治疗是生物瓣有使用期限的，目前大约十年都不到！所以对年轻人考虑机械瓣外科手术

郑重提示：线上咨询不能代替面诊，医生建议仅供参考！

加载更多

请编写代码，完成该项任务。

## Lecture2 Challenge

### Tips:

1.由于需要点击“加载更多”按钮，因此我们需要和网页进行交互。requests库无法解决这一问题，需要我們使用selenium模块，驱动浏览器打开网页，真正地点击按钮，完成交互。

selenium模块的使用教程：[https://blog.csdn.net/weixin\\_36279318/article/details/79475388](https://blog.csdn.net/weixin_36279318/article/details/79475388)

2.我们需要定位多个页面元素（因为要抓取多句对话），那么，如何使用xpath，在selenium中定位多个元素？

xpath的进阶教程：<https://cuiqingcai.com/2621.html>

xpath+selenium的一个简短实例：<https://www.cnblogs.com/liuliu-word/p/8057998.html>

3.如何将所抓取到的对话信息，存储到excel表格中？

我们建议使用xlwt模块，尝试自行百度寻找该模块的使用方法。

### Answer:

我会在10.6公布详细的注释代码。