**蒙古国新闻网（**[**https://gogo.mn**](https://gogo.mn)**）爬虫**

姓名：吴修宇

学号：2001210467

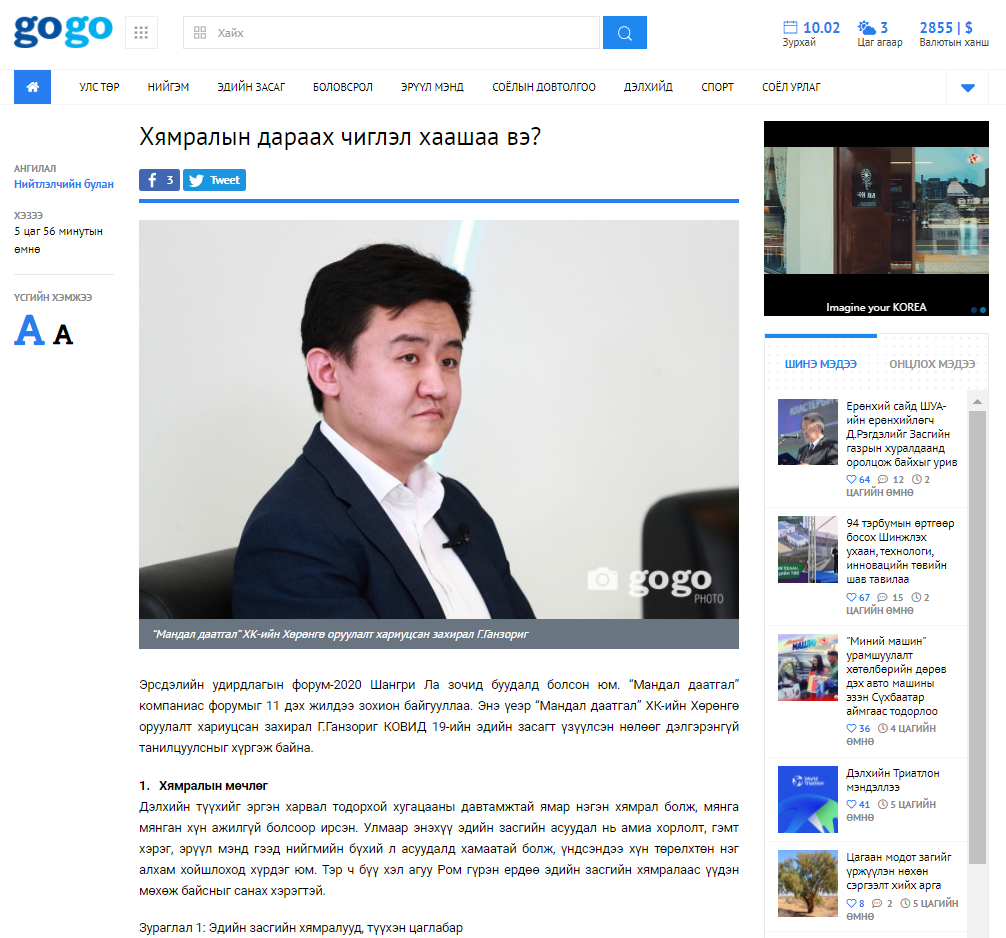
Github地址：https://github.com/flandy2010/spider-for-mongolian-news

## 一、需求

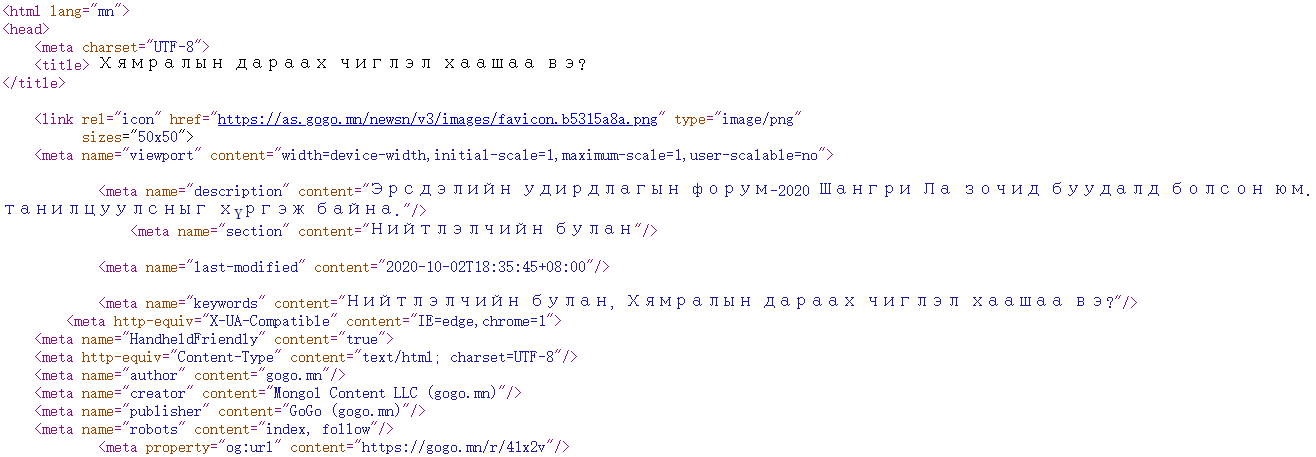
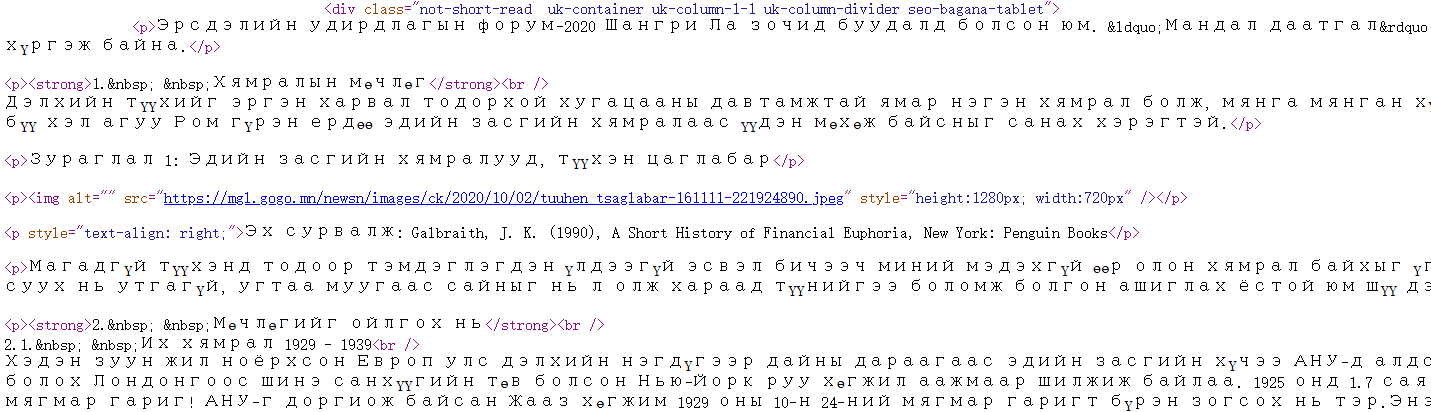
外国的新闻报道语法上较为规范，同时内容上会涵盖国家的重要事件。这样的文本既可以用于训练语言模型，又可以用于舆情分析、政治研究等。而gogo.mn蒙古新闻网是蒙古国四大新闻网站中更新最频繁，内容涵盖最广泛的。因此设计了一个针对该新闻网站的爬虫。

## 二、页面处理

2.1 页面格式



能从浏览器中看到的主要有标题和新闻内容两部分。通过查看网页源代码，可以看到，网页源码中还带有title，description，last-modified，keywords，section等信息，以及新闻内容的全文。



2.2 处理方法

对于网页的爬取主要使用requests库，处理上主要使用BeautifulSoup库以及正则表示式。

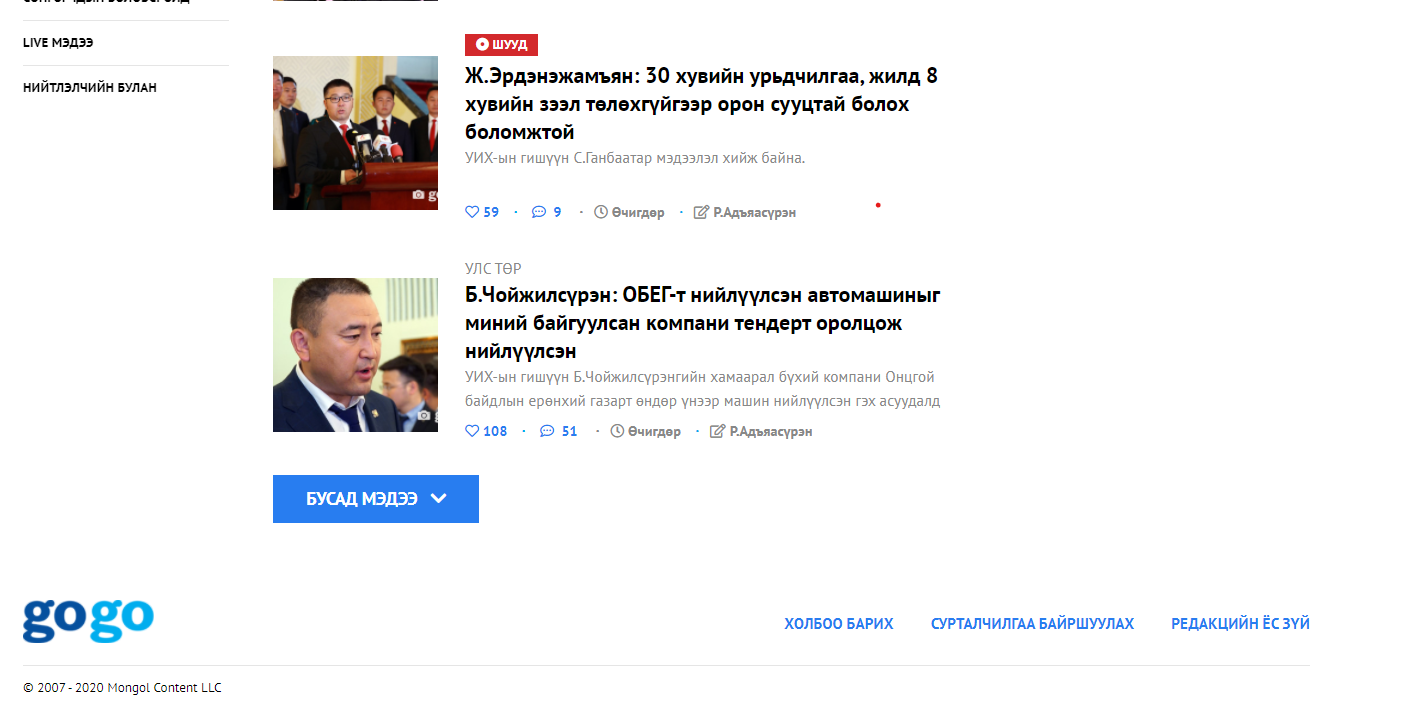
首先利用requests.get函数获取网页内容。然后利用BeautifulSoup库处理网页的内容。对于title，description，last-modified，keywords，section等，搜索<meta>标签，然后根据标签内的name属性的值，判断内容是什么。

对于新闻的正文部分，发现网站源码里面存在了两份数据。观察后发现，两份数据分属于不同的class下面。因此根据class属性选取其中一份。完成定位后，发现文字部分存在大量的格式，例如<strong></strong>，<span></span>，以及图片。考虑到无法枚举所有情况，因此将数据转化为字符串，利用正则表达式<.\*?>非贪婪匹配删去“<”和“>”，以及“<”“>”之间的内容，将剩下的作为新闻内容。

## 三、URL获取

每一条新闻拥有一个单独的URL。对于新闻URL的获取，主要利用的是selenium库以及BeautifulSoup库。

 观察到每一个新闻的URL的格式都是：<https://gogo.mn/r/>+五位编码，例如：<https://gogo.mn/r/41den>。观察到政治类新闻首页（<https://gogo.mn/i/2>）的源码中，带有该页中所显示的新闻的五位编码。但只会加载部分的新闻，想查看更多的新闻需要点击“БУСАД МЭДЭЭ（更多新闻）”按钮。



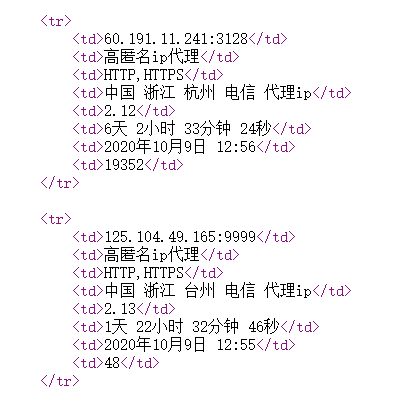
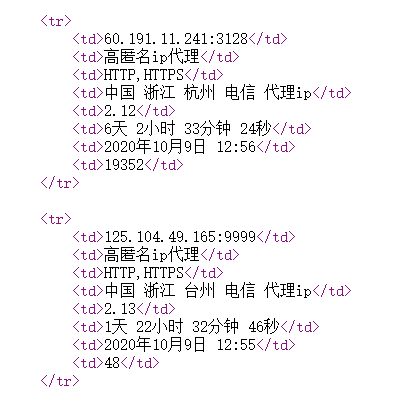
这时候，利用selenium库模拟浏览器的行为，根据按钮的xpath查到到按钮的位置，并且模拟点击。随后获取展开多次后网页的源代码，检索其中的标签a，并利用正则表达式判断href属性的值是不是/r/+五位数编码。

## 四、伪装机制

伪装机制主要使用了header伪装，ip伪装和控制爬取速度。

header伪装主要通过设置User-Agent和Referer。将一个爬虫伪装成一个从[www.gogo.mn](http://www.gogo.mn)网站点击链接进来看新闻的浏览器。

 ip伪装是设置爬虫的ip。由于没钱购买代理池，选择了从免费ip代理网站爬取ip – 筛选可用ip – 建立ip池的方式。首先选择了一个提供免费ip代理的网站：<http://www.xiladaili.com/>，将网站提供的ip爬下来。



随后使用该ip代理访问https://httpbin.org/ip，通过返回值判断该ip代理是否可用。如果可用则加入ip池中。

在爬取过程中，会控制爬取速度，每爬取一张网页会随机休眠2-5秒。如果爬取失败，休眠时间会随着连续失败次数增加而限性增长。

## 五、样例

## 六、存在的问题

1、在模拟点击“更多新闻”按钮的时候，偶尔会出现程序卡住的情况。重新运行程序则会正常运行。暂时没有找到可复现的方式。

2、从免费ip代理网站上爬取的代理，对于蒙古国新闻网的可用性不高。