汽车之家数据采集说明

陈肖

1. 爬虫策略说明

本次爬虫的入口是从<http://www.autohome.com.cn/car/?pvareaid=101452> 进入，此网页是动态加载的，由于暂时无法解动态加载后的解码问题（一部分代码不是用gbk形式加载，导致一直出问题，用了一天时间，仍然没有成功解决这个问题）。之后，迫于无奈，改变了策略。我用的方法是将整个网页保存下来，再进行读取。网页可见文件中的data.html 以及data.py。之后按照需要的层次进行爬取。

1. 采集的数据

采集的数据以csv格式保存。本次爬取的数据只爬取了部分表格中的，从安全装备之后的项目，没有进行爬取。（如果对格式有需要或者对个别项目有需要，可以告诉我）。未爬取的项目如图1所示



图1

1. 部分未采集到数据

对于根目录网页，如图2所示



图2

我一开始以为只要是灰色的条目，就没有参数配置。但最后检查的时候发现，有极个别（30条以内）的条目，即使灰色任然在二手车部分有参数配置。由于一部分这些网页的网页结构和之前的网页不同，而另一部分相同但是没有参数配置。导致编写爬虫的路径比较困难，对于这一小部分数据，并没有能够成功采集。