知乎

控制字符概述-以及移除

Dytttf
人生苦短 只用python

背景

经常做爬虫的人，应该对\x00、\x01这样的字符不陌生，网页源码里面 不经常 出现。不过一般都不深究这到底是啥。一开始我也没研究，发现之后就拿正则替换掉，简单粗暴的处理。之所以要去掉，是因为使用Python的lxml库处理的时候会抛异常。再后来，由于需要做一个通用一些的采集器，没办法再无视这个问题了，于是Google一番之后，Copy了一段代码：

```
text = re.sub("[\x00-\x1F\x7F]\s*", "", text)
```

不错，测试下来运行良好。完美去掉了所有控制字符。只是好景不长，过了几个月，在排查一个Xpath解析问题的时候，突然发现这段代码有很大问题。。。

控制字符简述

以下内容抄自维基百科：zh.wikipedia.org/wiki/%...

> 控制字符，是出现在特定的信息文本中，表示某一控制功能的字符。

好吧，这个东西最开始设计是为了给打印机用的。。。 ，现在的话，很多控制字符感觉都没啥用了。简单举几个例子：

- 键盘上的 Backspace 产生 \x08, 用来回退一格，或者删除前一个字符
- Return 或 Enter健产生 \x10、\x13，用来换行。说人话就是 \n\r
- Tab健产生 \x09,也就是 \t

其他的就看Wiki吧，写的挺清楚的。

遇到的问题以及解决方案

例子中的这几个之所以列出，是因为我遇到的问题就是他们几个引起的。一般的网页源码中，存在的\t、\r、\n等字符，都是用来调整样式的，对数据没啥影响。但是今天就碰到了很坑的网站。它是这么写源码的：

知乎

卧槽。竟然用\n而不是空格来分割标签名与属性。（顺便说一句。空格是\x20，但不是控制字符）

于是。我抄的正则就有问题了。把\n给替换掉了之后，源码就变成了

```
<divclass='content'>.....</div>
```

这还怎么解析。。。

于是修改正则如下：

```
text = re.sub("[\x00-\x08\x11\x12\x14-\x1F\x7F\x80-\x9F]", "", text)
```

跳过了这几个常用字符。完美收工。 感觉应该、也许、大概、可能不会再出问题了吧！

后记

果然还是出问题了。。。

正则写顺手了。。。 应该是忽略16进制的9、10、13的，也就是\x09、\x10、\x0D。

```
text = re.sub("[\x00-\x08\x0B\x0C\x0E-\x1F\x7F-\x9F]", "", text)
```

编辑于 2020-07-05

字符编码　Unicode（统一码）　字符

推荐阅读

你需要知道的字符串编码

在打开网页或者文件的时候，你…